**IN THE UNITED STATES PATENT AND TRADEMARK OFFICE**

| | |
|---|---|
| Attorney Docket No.: | **DEX-0172** |
| Inventors: | **Salceda et al.** |
| Serial No.: | **09/763,978** |
| Filing Date: | **April 25, 2001** |
| Examiner: | **Helms, Larry Ronald** |
| Customer No.: | **32800** |
| Group Art Unit: | **1642** |
| Confirmation No.: | **6964** |
| Title: | **A Novel Method of Diagnosing, Monitoring, Staging, Imaging and Treating Various Cancers** |



**Declaration by Dr. Susana Salceda**

I, Susana Salceda, hereby declare:

1. I was awarded a Masters of Science in Biochemistry in 1983 and a Ph.D. in Biochemistry in 1990, both from the School of Science at the University of Buenos Aires, Argentina. After obtaining my Ph.D., I served as a postdoctoral researcher at Thomas Jefferson University from 1991 to 1998. While at Thomas Jefferson University I contributed to the analysis of mechanisms of oxygen sensing, signal transduction and regulation of gene expression by hypoxia and other stimuli.

From 1998 to 2002, I worked in the Gene Discovery division at diaDexus, Inc. holding the position of

Scientist. At diaDexus I contributed to research using genomics based analyses focusing on the discovery, identification and characterization of novel polynucleotides and encoded proteins differentially expressed in cancer. Identified polynucleotides and encoded proteins were used to develop novel diagnostic and therapeutic products for the improved detection, classification, prognosis and treatment of cancer.

Since 2002, I have been a Senior Scientist working in the Expression Product Development Department at Affymetrix, Inc., in Santa Clara, CA. At Affymetrix I contribute to the development of new assays and reagents to process DNA and RNA samples for microarray analysis.

2. As a scientist, a former diaDexus employee, and a named inventor, I am familiar with the teachings of the above-referenced patent application. I was responsible for the discovery of Ovr110 and the sequences encoding it.

3. I have reviewed and am familiar with the office action in the above-referenced patent application dated June 22, 2005 from the U.S. Patent Office.

4. I understand the Examiner has taken a position that the "invention is not supported by either a substantial asserted utility or a well established utility." I respectfully disagree.

5. At the time of the invention the usefulness of an isolated antibody or antibody fragment that binds specifically to a cancer marker such as the protein encoded by polynucleotide SEQ ID NO: 1 was well known.

6. Further, at the time of the invention we routinely obtained a protein sequence or open reading frame from information related to a polynucleotide sequence such as that provided for the polynucleotide sequence of SEQ ID NO: 1.

For example, as shown in Examples 1 and 2 of the above referenced patent application, the sequence and expression data of SEQ ID NO:1 is based on an mRNA molecule and therefore has a set 5′ to 3′ orientation. Thus, from this information, we know the protein is encoded in the forward (5′ to 3′) direction of SEQ ID NO: 1.

Furthermore, since expressed mRNA encode for proteins we know that the open reading frame in the forward direction of SEQ ID NO: 1 would be in a frame encoding for a Methionine near the 5′ end, encode many amino acids and terminate with a stop codon. Thus, any reading frame sequence of SEQ ID NO: 1 with lots of stop codons can be ruled out since we know to look for a long open reading frame sequence beginning with an M and ending with a stop codon in accordance with the information taught in the patent application about SEQ ID NO: 1.

By 1998 there were many tools available for use to determine either the protein sequence or the open reading frame (ORF) of a sequence such as SEQ ID NO: 1. Examples of such programs include the MAP[1] application, part of the GCG software suite from Accelrys Software Inc. (San Diego, CA), the Translate application, part of ExPASy (Expert Protein Analysis System) available online (at www.expasy.org/tools/dna.html) from the Swiss Institute of

[1] Devereux J, Haeberli P, Smithies O. (1984 NAR 11, 387-395)

Bioinformatics (Lausanne, Switzerland) and the ORF Finder (Open Reading Frame Finder) application available online (at www.ncbi.nlm.nih.gov/gorf/gorf.html)from the National Center for Biotechnology Information (NCBI) (Bethesda, MD).

As examples, attached are the results of the MAP, Translate and ORF Finder programs described above. The attached MAP program results (Figure 1) display SEQ ID NO: 1 as taught in the patent application in the forward direction, the reverse complement strand, and the protein translation of the three frames of the forward nucleotide strand followed by the protein translation of the three frames of the reverse compliment strand. For clarity, the open reading frame and protein encoded by SEQ ID NO: 1 have been underlined. As with many programs, the start codons encoding a Methionine (denoted by "**M**" or "**Met**") and stop codons not encoding an amino acid (denoted by "*" or "**Stop**") are in bold. Also displayed in the MAP results, but not relevant to the open reading frame or encoded protein, are the nucleotide restriction sites for the endonuclease SAU3AI.

The attached Translate program results (Figure 2) display the protein translations of the three forward frames (5'3') followed by the protein translation of the three frames of the reverse compliment strand (3'5'). For clarity, the protein encoded by SEQ ID NO: 1 has been underlined.

The attached ORF Finder program results (Figure 3) displays a graphical representation of the ORFs greater than 100 nucleotides in length in each of the six frames of SEQ ID NO: 1. The longest open reading frame is listed first on the right as frame +2 from nucleotide 62-910 with a length of 849 nucleotides. This open reading frame is

selected (highlighted) in the display and the ORF nucleotide sequence and encoded 282 amino acid protein sequence is displayed below.

Using the attached results from the MAP application, Translate application, ORF Finder application, or output from another simple translation program, the encoded protein and open reading frame are clear. Here MAP, Translate or ORF Finder show the protein encoded by SEQ ID NO: 1 is 282 amino acids long. Thus, using only the information taught in the specification as filed, the open reading frame for SEQ ID NO: 1 and the encoded protein can be routinely and unambiguously identified.

7. The Examiner also suggests that there was "no indication of what the protein [encoded by SEQ ID NO: 1] was." I respectfully disagree. As shown by the attached results from the MAP application, Translate application and ORF Finder application, the protein encoded by SEQ ID NO: 1 was readily obtainable with tools used routinely as of 1998.

8. Similarly, the process of expressing the protein encoded by a nucleotide such as SEQ ID NO: 1 and generating antibodies to the protein was well known as of 1998 and prior thereto.

9. I respectfully disagree with the Examiner's suggestion that this sequence and invention are "starting points for further research and investigation into potential practical uses." As shown herein, the nucleotide sequence of SEQ ID NO: 1 and the characteristics disclosed in the patent application about SEQ ID NO: 1 were adequate

to routinely and unambiguously obtain the protein sequence and then generate antibodies or antibody fragments thereto.

10. I also respectfully disagree with the Examiner's suggestions that "one would not have known a utility for such a protein" and that the "specification does not teach a utility for use of the antibody." The patent application teaches that "the mRNA overexpression in most of the matching samples tested are indicative of Ovr110... being a diagnostic marker for gynecologic cancers." Further, uses for the protein expressed by the CSG encoded by SEQ ID NO: 1 are explicitly described in the specification. Since the mRNA of SEQ ID NO: 1 is overexpressed in gynecologic cancers samples, and encodes a protein, the value of antibodies to this protein to detect overexpressed protein in gynecologic cancers would also be understood.

Further, the specification explicitly teaches that antibodies against Cancer Specific Genes (CSG) such as SEQ ID NO: 1 "can be used to detect or image localization of CSG in a patient for the purpose of detecting or diagnosing selected cancers."

The specification also explicitly teaches that antibodies against Cancer Specific Genes (CSG) such as SEQ ID NO: 1 "can be injected into a patient suspected of having a selected cancer for diagnostic and/or therapeutic purposes."

Furthermore, contrary to the Examiner's suggestion, the specification provides detailed teachings as to how one of skill in the art could use these antibodies in an ELISA assay or a competition assay to detect cancer, thus providing guidance regarding use of the invention "in a manner that constitutes a substantial utility."

11. I also respectfully disagree with the Examiner's suggestion that "applicants were not in possession of any protein encoded by SEQ ID NO: 1." As I showed herein, using standard tools available at the time of the invention, one of skill in the art could readily determine the protein encoded by SEQ ID NO: 1. All the necessary information to do so is provided by the polynucleotide sequence and the characteristics of this sequence taught in the patent application.

*I hereby declare that all statements herein of my own knowledge are true and that all statements made on information or belief are believed to be true; and further that these statements were made with the knowledge that willful statements and the like so made are punishable by fine or by imprisonment, or both, under §1001 of Title 18 of the United States code, and that such willful statements may jeopardize the validity of the application, any patent issuing there upon, or any patent to which this verified statement is directed.*

Susana Salceda, Ph.D.

10/18/05
Date

# FIGURE 1

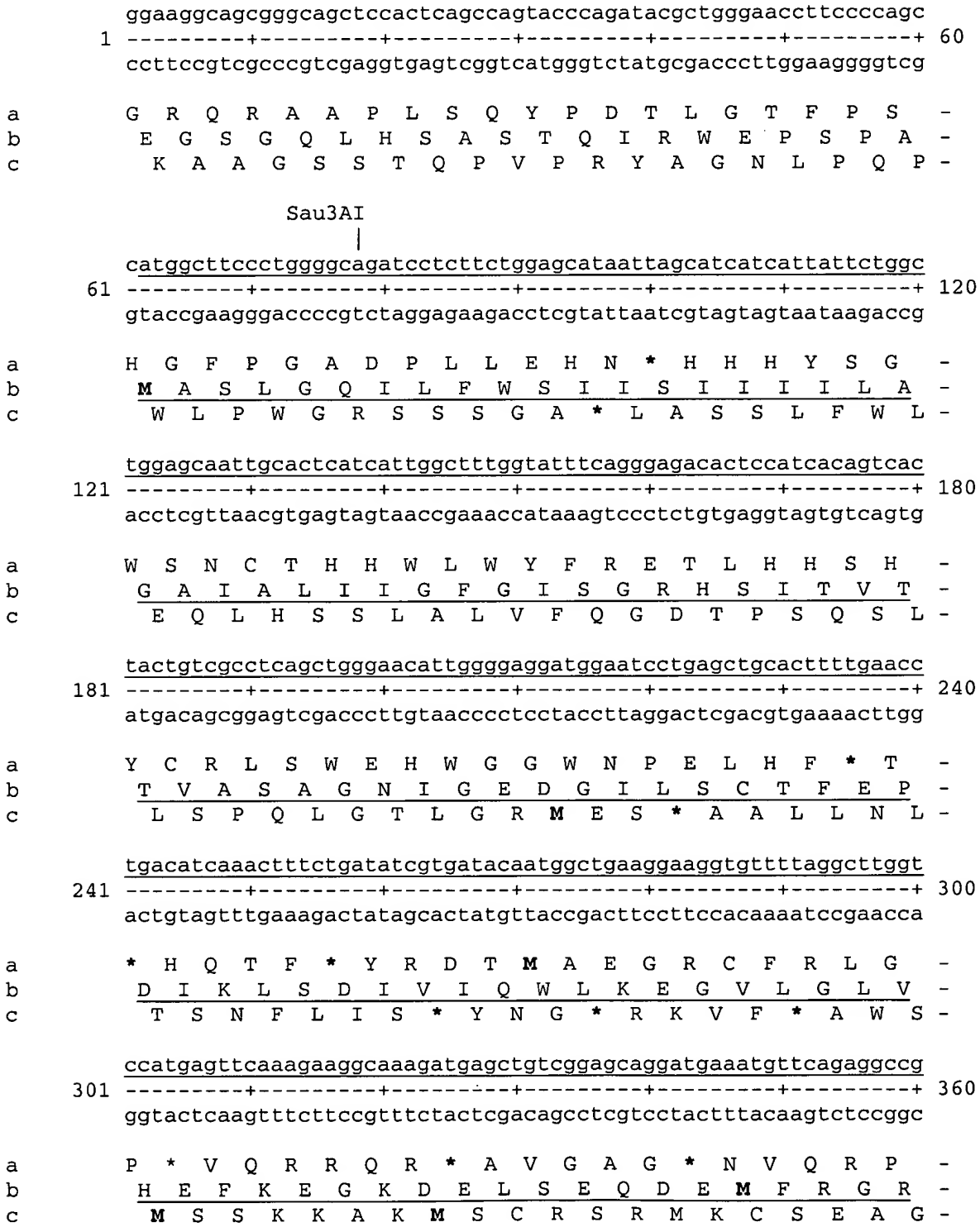


(Linear) MAP of: dex0043_1.seq check: 5695 from: 1 to: 2587

DEX0043_1

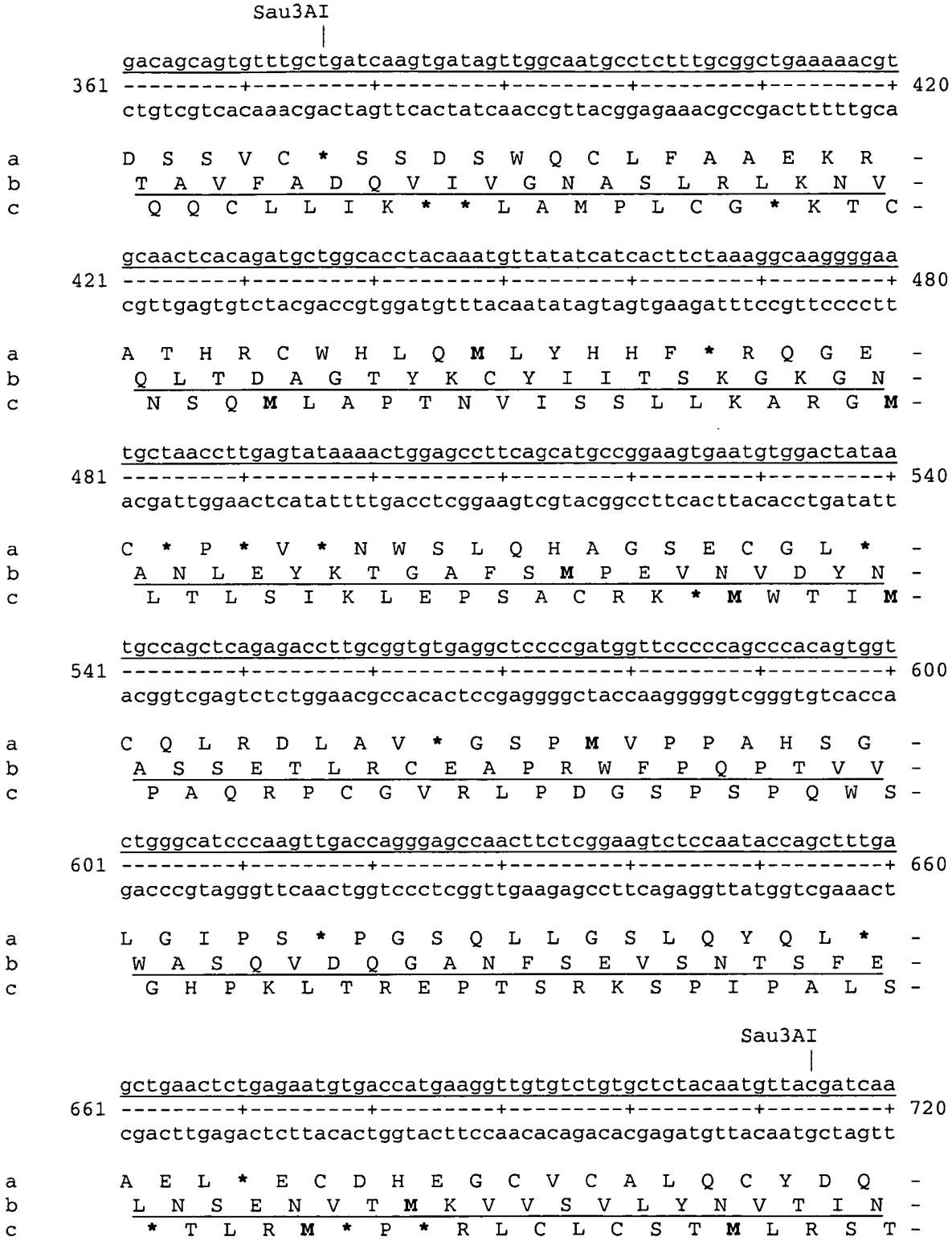
With 1 enzymes: SAU3AI

Forward frame translations:



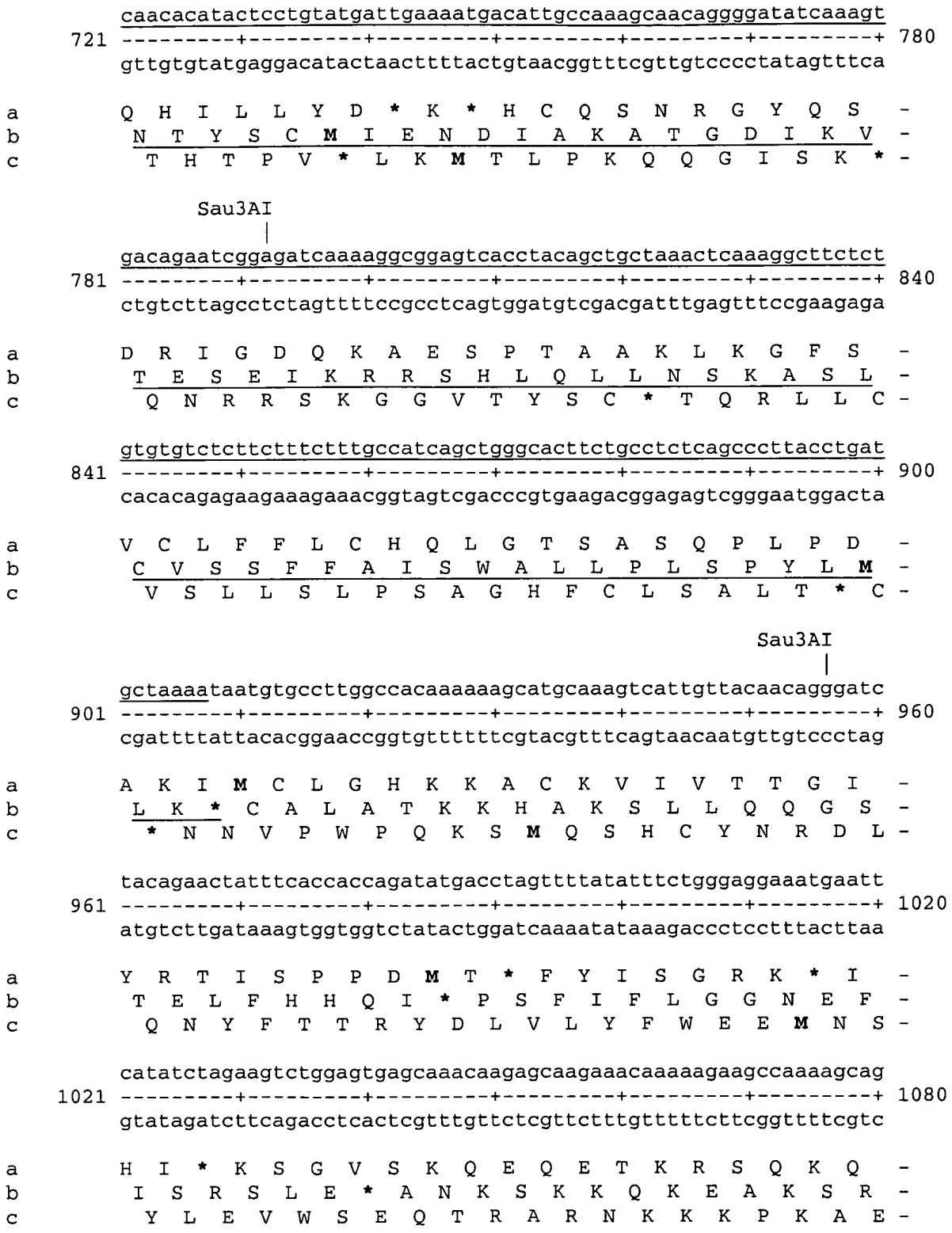
# FIGURE 1




Sau3AI
gacagcagtgtttgctgatcaagtgatagttggcaatgcctctttgcggctgaaaaacgt
361
ctgtcgtcacaaacgactagttcactatcaaccgttacggagaaacgccgacttttgca
420
a D S S V C * S S D S W Q C L F A A E K R -
b T A V F A D Q V I V G N A S L R L K N V -
c Q Q C L L I K * * L A M P L C G * K T C -
gcaactcacagatgctggcacctacaaatgttatatcatcacttctaaaggcaaggggaa
421
cgttgagtgtctacgaccgtggatgtttacaatatagtagtgaagatttccgttcccctt
480
a A T H R C W H L Q M L Y H H F * R Q G E -
b Q L T D A G T Y K C Y I I T S K G K G N -
c N S Q M L A P T N V I S S L L K A R G M -
tgctaaccttgagtataaaactggagccttcagcatgccggaagtgaatgtggactataa
481
acgattggaactcatattttgacctcggaagtcgtacggccttcacttacacctgatatt
540
a C * P * V * N W S L Q H A G S E C G L * -
b A N L E Y K T G A F S M P E V N V D Y N -
c L T L S I K L E P S A C R K * M W T I M -
tgccagctcagagaccttgcggtgtgaggctccccgatggttcccccagcccacagtggt
541
acggtcgagtctctggaacgccacactccgaggggctaccaaggggtcgggtgtcacca
600
a C Q L R D L A V * G S P M V P P A H S G -
b A S S E T L R C E A P R W F P Q P T V V -
c P A Q R P C G V R L P D G S P S P Q W S -
ctgggcatcccaagttgaccagggagccaacttctcggaagtctccaataccagctttga
601
gacccgtagggttcaactggtccctcggttgaagagccttcagaggttatggtcgaaact
660
a L G I P S * P G S Q L L G S L Q Y Q L * -
b W A S Q V D Q G A N F S E V S N T S F E -
c G H P K L T R E P T S R K S P I P A L S -
Sau3AI
gctgaactctgagaatgtgaccatgaaggttgtgtctgtgctctacaatgttacgatcaa
661
cgacttgagactcttacactggtacttccaacacagacacgagatgttacaatgctagtt
720
a A E L * E C D H E G C V C A L Q C Y D Q -
b L N S E N V T M K V V S V L Y N V T I N -
c * T L R M * P * R L C L C S T M L R S T -

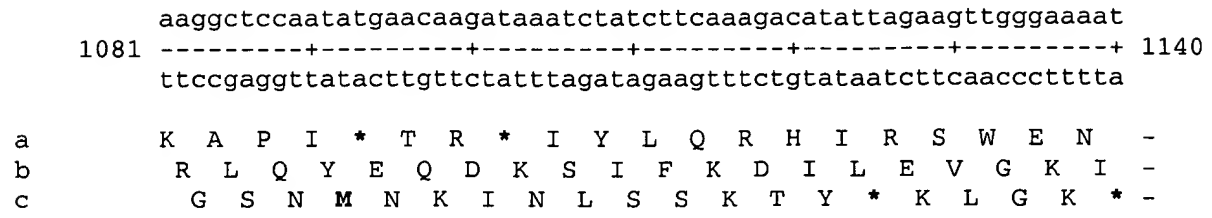
# FIGURE 1




caacacatactcctgtatgattgaaaatgacattgccaaagcaacaggggatatcaaagt
721 780
gttgtgtatgaggacatactaacttttactgtaacggtttcgttgtcccctatagtttca
a Q H I L L Y D * K * H C Q S N R G Y Q S -
b N T Y S C M I E N D I A K A T G D I K V -
c T H T P V * L K M T L P K Q Q G I S K * -
Sau3AI
gacagaatcggagatcaaaaggcggagtcacctacagctgctaaactcaaaggcttctct
781 840
ctgtcttagcctctagttttccgcctcagtggatgtcgacgatttgagtttccgaagaga
a D R I G D Q K A E S P T A A K L K G F S -
b T E S E I K R R S H L Q L L N S K A S L -
c Q N R R S K G G V T Y S C * T Q R L L C -
gtgtgtctcttctttctttgccatcagctgggcacttctgcctctcagcccttacctgat
841 900
cacacagagaagaaagaaacggtagtcgacccgtgaagacggagagtcgggaatggacta
a V C L F F L C H Q L G T S A S Q P L P D -
b C V S S F F A I S W A L L P L S P Y L M -
c V S L L S L P S A G H F C L S A L T * C -
Sau3AI
gctaaaataatgtgccttggccacaaaaaagcatgcaaagtcattgttacaacagggatc
901 960
cgattttattacacggaaccggtgtttttttcgtacgtttcagtaacaatgttgtccctag
a A K I M C L G H K K A C K V I V T T G I -
b L K * C A L A T K K H A K S L L Q Q G S -
c * N N V P W P Q K S M Q S H C Y N R D L -
tacagaactatttcaccaccagatatgacctagttttatatttctgggaggaaatgaatt
961 1020
atgtcttgataaagtggtggtctatactggatcaaaatataaagaccctcctttacttaa
a Y R T I S P P D M T * F Y I S G R K * I -
b T E L F H H Q I * P S F I F L G G N E F -
c Q N Y F T T R Y D L V L Y F W E E M N S -
catatctagaagtctggagtgagcaaacaagagcaagaaacaaaaagaagccaaaagcag
1021 1080
gtatagatcttcagacctcactcgtttgttctcgttctttgtttttcttcggttttcgtc
a H I * K S G V S K Q E Q E T K R S Q K Q -
b I S R S L E * A N K S K K Q K E A K S R -
c Y L E V W S E Q T R A R N K K K P K A E -

# FIGURE 1




aaggctccaatatgaacaagataaatctatcttcaaagacatattagaagttgggaaaat
1081 ---------+---------+---------+---------+---------+---------+ 1140
ttccgaggttatacttgttctatttagatagaagtttctgtataatcttcaacccttta
a K A P I * T R * I Y L Q R H I R S W E N -
b R L Q Y E Q D K S I F K D I L E V G K I -
c G S N M N K I N L S S K T Y * K L G K * -

# FIGURE 1

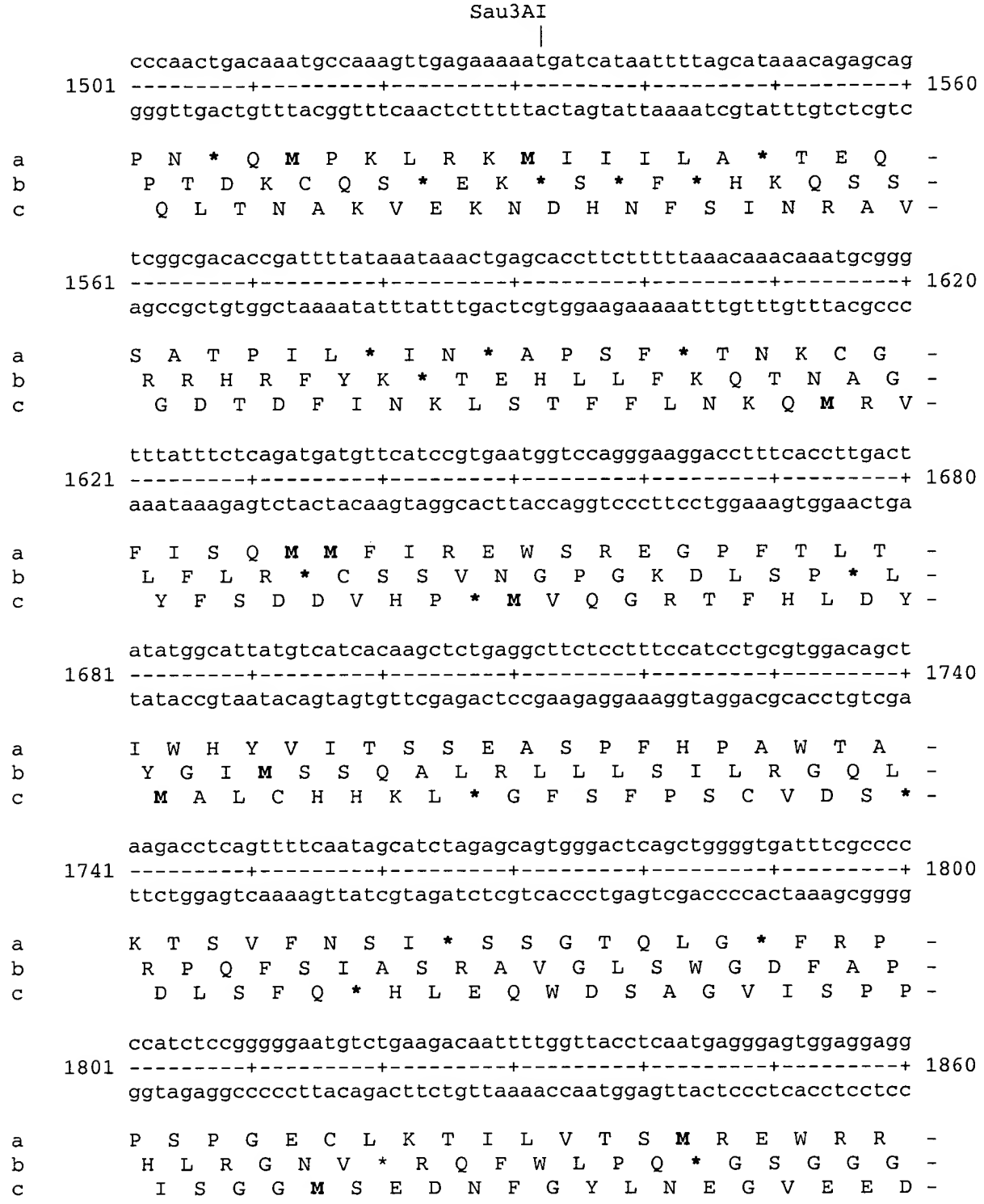


Reverse frame translations:

```
      aattcatgtgaactagacaagtgtgttaagagtgataagtaaaatgcacgtggagacaag
 1141 ---------+---------+---------+---------+---------+---------+ 1200
      ttaagtacacttgatctgttcacacaattctcactattcattttacgtgcacctctgttc

a       N  S  C  E  L  D  K  C  V  K  S  D  K  *  N  A  R  G  D  K  -
b        I  H  V  N  *  T  S  V  L  R  V  I  S  K  M  H  V  E  T  S -
c         F  M  *  T  R  Q  V  C  *  E  *  *  V  K  C  T  W  R  Q  V -

            Sau3AI
              |
      tgcatccccagatctcagggacctccccctgcctgtcacctggggagtgagaggacagga
 1201 ---------+---------+---------+---------+---------+---------+ 1260
      acgtaggggtctagagtccctggaggggggacggacagtggacccctcactctcctgtcct

a       C  I  P  R  S  Q  G  P  P  P  A  C  H  L  G  S  E  R  T  G  -
b        A  S  P  D  L  R  D  L  P  L  P  V  T  W  G  V  R  G  Q  D -
c         H  P  Q  I  S  G  T  S  P  C  L  S  P  G  E  *  E  D  R  I -

      tagtgcatgttctttgtctctgaatttttagttatatgtgctgtaatgttgctctgagga
 1261 ---------+---------+---------+---------+---------+---------+ 1320
      atcacgtacaagaaacagagacttaaaaatcaatatacacgacattacaacgagactcct

a       *  C  M  F  F  V  S  E  F  L  V  I  C  A  V  M  L  L  *  G  -
b        S  A  C  S  L  S  L  N  F  *  L  Y  V  L  *  C  C  S  E  E -
c         V  H  V  L  C  L  *  I  F  S  Y  M  C  C  N  V  A  L  R  K -

      agcccctggaaagtctatcccaacatatccacatcttatattccacaaattaagctgtag
 1321 ---------+---------+---------+---------+---------+---------+ 1380
      tcggggacctttcagatagggttgtataggtgtagaatataaggtgtttaattcgacatc

a       S  P  W  K  V  Y  P  N  I  S  T  S  Y  I  P  Q  I  K  L  *  -
b        A  P  G  K  S  I  P  T  Y  P  H  L  I  F  H  K  L  S  C  S -
c         P  L  E  S  L  S  Q  H  I  H  I  L  Y  S  T  N  *  A  V  V -

      tatgtaccctaagacgctgctaattgactgccacttcgcaactcaggggcggctgcattt
 1381 ---------+---------+---------+---------+---------+---------+ 1440
      atacatgggattctgcgacgattaactgacggtgaagcgttgagtccccgccgacgtaaa

a       Y  V  P  *  D  A  A  N  *  L  P  L  R  N  S  G  A  A  A  F  -
b        M  Y  P  K  T  L  L  I  D  C  H  F  A  T  Q  G  R  L  H  F -
c         C  T  L  R  R  C  *  L  T  A  T  S  Q  L  R  G  G  C  I  L -

      tagtaatgggtcaaatgattcacttttatgatgcttccaaaggtgccttggcttctctt
 1441 ---------+---------+---------+---------+---------+---------+ 1500
      atcattacccagtttactaagtgaaaaatactacgaaggtttccacggaaccgaagagaa

a       *  *  W  V  K  *  F  T  F  Y  D  A  S  K  G  A  L  A  S  L  -
b        S  N  G  S  N  D  S  L  F  M  M  L  P  K  V  P  W  L  L  F -
c         V  M  G  Q  M  I  H  F  L  *  C  F  Q  R  C  L  G  F  S  S -
```

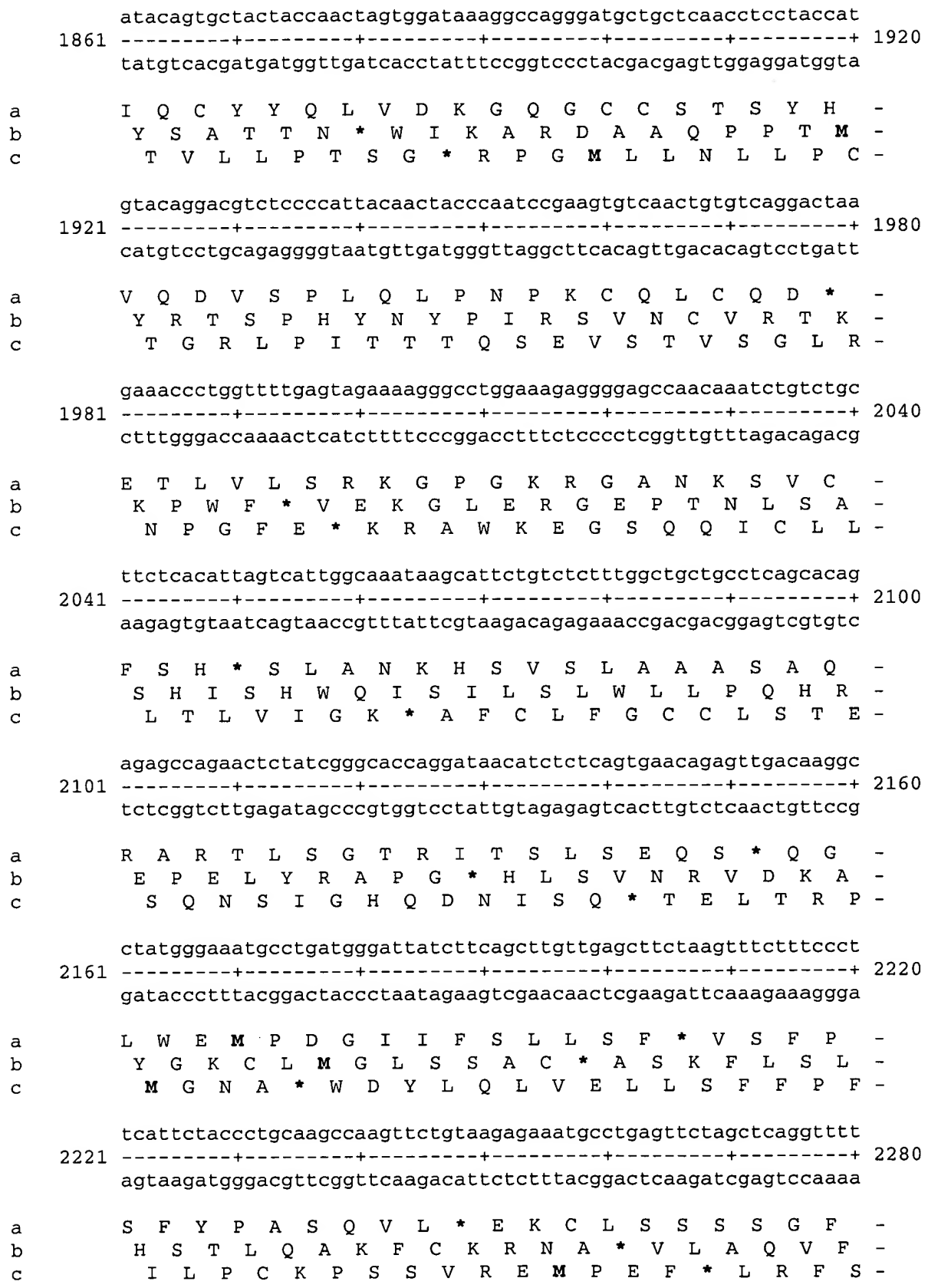
# FIGURE 1




Sau3AI
cccaactgacaaatgccaaagttgagaaaaatgatcataattttagcataaacagagcag
1501 ---------+---------+---------+---------+---------+---------+ 1560
gggttgactgtttacggtttcaactctttttactagtattaaaatcgtatttgtctcgtc
a P N * Q M P K L R K M I I I L A * T E Q -
b P T D K C Q S * E K * S * F * H K Q S S -
c Q L T N A K V E K N D H N F S I N R A V -
tcggcgacaccgattttataaataaactgagcaccttctttttaaacaaacaaatgcggg
1561 ---------+---------+---------+---------+---------+---------+ 1620
agccgctgtggctaaaatatttatttgactcgtggaagaaaaatttgtttgtttacgccc
a S A T P I L * I N * A P S F * T N K C G -
b R R H R F Y K * T E H L L F K Q T N A G -
c G D T D F I N K L S T F F L N K Q M R V -
tttatttctcagatgatgttcatccgtgaatggtccagggaaggacctttcaccttgact
1621 ---------+---------+---------+---------+---------+---------+ 1680
aaataaagagtctactacaagtaggcacttaccaggtcccttcctggaaagtggaactga
a F I S Q M M F I R E W S R E G P F T L T -
b L F L R * C S S V N G P G K D L S P * L -
c Y F S D D V H P * M V Q G R T F H L D Y -
atatggcattatgtcatcacaagctctgaggcttctcctttccatcctgcgtggacagct
1681 ---------+---------+---------+---------+---------+---------+ 1740
tataccgtaatacagtagtgttcgagactccgaagaggaaaggtaggacgcacctgtcga
a I W H Y V I T S S E A S P F H P A W T A -
b Y G I M S S Q A L R L L L S I L R G Q L -
c M A L C H H K L * G F S F P S C V D S * -
aagacctcagttttcaatagcatctagagcagtgggactcagctggggtgatttcgcccc
1741 ---------+---------+---------+---------+---------+---------+ 1800
ttctggagtcaaaagttatcgtagatctcgtcaccctgagtcgaccccactaaagcgggg
a K T S V F N S I * S S G T Q L G * F R P -
b R P Q F S I A S R A V G L S W G D F A P -
c D L S F Q * H L E Q W D S A G V I S P P -
ccatctccgggggaatgtctgaagacaattttggttacctcaatgagggagtggaggagg
1801 ---------+---------+---------+---------+---------+---------+ 1860
ggtagaggcccccttacagacttctgttaaaaccaatggagttactccctcacctcctcc
a P S P G E C L K T I L V T S M R E W R R -
b H L R G N V * R Q F W L P Q * G S G G G -
c I S G G M S E D N F G Y L N E G V E E D -

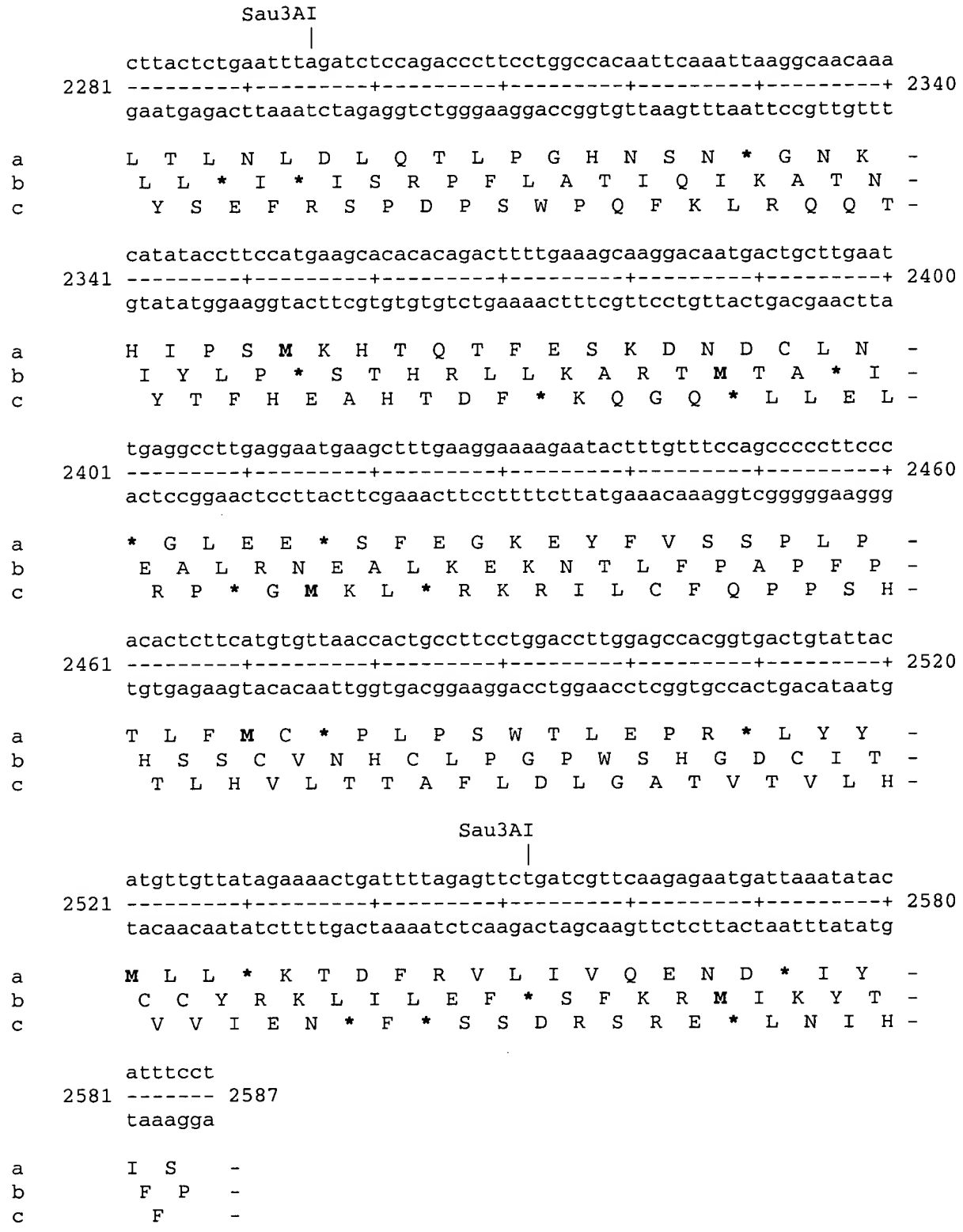
# FIGURE 1




atacagtgctactaccaactagtggataaaggccagggatgctgctcaacctcctaccat
1861 ---------+---------+---------+---------+---------+---------+ 1920
tatgtcacgatgatggttgatcacctatttccggtccctacgacgagttggaggatggta
a I Q C Y Y Q L V D K G Q G C C S T S Y H -
b Y S A T T N * W I K A R D A A Q P P T M -
c T V L L P T S G * R P G M L L N L L P C -
gtacaggacgtctccccattacaactacccaatccgaagtgtcaactgtgtcaggactaa
1921 ---------+---------+---------+---------+---------+---------+ 1980
catgtcctgcagaggggtaatgttgatgggttaggcttcacagttgacacagtcctgatt
a V Q D V S P L Q L P N P K C Q L C Q D * -
b Y R T S P H Y N Y P I R S V N C V R T K -
c T G R L P I T T T Q S E V S T V S G L R -
gaaaccctggttttgagtagaaaagggcctggaaagaggggagccaacaaatctgtctgc
1981 ---------+---------+---------+---------+---------+---------+ 2040
ctttgggaccaaaactcatcttttcccggacctttctcccctcggttgtttagacagacg
a E T L V L S R K G P G K R G A N K S V C -
b K P W F * V E K G L E R G E P T N L S A -
c N P G F E * K R A W K E G S Q Q I C L L -
ttctcacattagtcattggcaaataagcattctgtctctttggctgctgcctcagcacag
2041 ---------+---------+---------+---------+---------+---------+ 2100
aagagtgtaatcagtaaccgtttattcgtaagacagagaaaccgacgacggagtcgtgtc
a F S H * S L A N K H S V S L A A A S A Q -
b S H I S H W Q I S I L S L W L L P Q H R -
c L T L V I G K * A F C L F G C C L S T E -
agagccagaactctatcgggcaccaggataacatctctcagtgaacagagttgacaaggc
2101 ---------+---------+---------+---------+---------+---------+ 2160
tctcggtcttgagatagcccgtggtcctattgtagagagtcacttgtctcaactgttccg
a R A R T L S G T R I T S L S E Q S * Q G -
b E P E L Y R A P G * H L S V N R V D K A -
c S Q N S I G H Q D N I S Q * T E L T R P -
ctatgggaaatgcctgatgggattatcttcagcttgttgagcttctaagtttctttccct
2161 ---------+---------+---------+---------+---------+---------+ 2220
gataccctttacggactaccctaatagaagtcgaacaactcgaagattcaaagaaaggga
a L W E M P D G I I F S L L S F * V S F P -
b Y G K C L M G L S S A C * A S K F L S L -
c M G N A * W D Y L Q L V E L L S F F P F -
tcattctaccctgcaagccaagttctgtaagagaaatgcctgagttctagctcaggtttt
2221 ---------+---------+---------+---------+---------+---------+ 2280
agtaagatgggacgttcggttcaagacattctctttacggactcaagatcgagtccaaaa
a S F Y P A S Q V L * E K C L S S S S G F -
b H S T L Q A K F C K R N A * V L A Q V F -
c I L P C K P S S V R E M P E F * L R F S -

# FIGURE 1




Sau3AI
cttactctgaatttagatctccagacccttcctggccacaattcaaattaaggcaacaaa
2281 ---------+---------+---------+---------+---------+---------+ 2340
gaatgagacttaaatctagaggtctgggaaggaccggtgttaagtttaattccgttgttt
a L T L N L D L Q T L P G H N S N * G N K -
b L L * I * I S R P F L A T I Q I K A T N -
c Y S E F R S P D P S W P Q F K L R Q Q T -
catataccttccatgaagcacacacagacttttgaaagcaaggacaatgactgcttgaat
2341 ---------+---------+---------+---------+---------+---------+ 2400
gtatatggaaggtacttcgtgtgtgtctgaaaactttcgttcctgttactgacgaactta
a H I P S M K H T Q T F E S K D N D C L N -
b I Y L P * S T H R L L K A R T M T A * I -
c Y T F H E A H T D F * K Q G Q * L L E L -
tgaggccttgaggaatgaagctttgaaggaaaagaatactttgtttccagccccttccc
2401 ---------+---------+---------+---------+---------+---------+ 2460
actccggaactccttacttcgaaacttccttttcttatgaaacaaaggtcgggggaaggg
a * G L E E * S F E G K E Y F V S S P L P -
b E A L R N E A L K E K N T L F P A P F P -
c R P * G M K L * R K R I L C F Q P P S H -
acactcttcatgtgttaaccactgccttcctggaccttggagccacggtgactgtattac
2461 ---------+---------+---------+---------+---------+---------+ 2520
tgtgagaagtacacaattggtgacggaaggacctggaacctcggtgccactgacataatg
a T L F M C * P L P S W T L E P R * L Y Y -
b H S S C V N H C L P G P W S H G D C I T -
c T L H V L T T A F L D L G A T V T V L H -
Sau3AI
atgttgttatagaaaactgattttagagttctgatcgttcaagagaatgattaaatatac
2521 ---------+---------+---------+---------+---------+---------+ 2580
tacaacaatatcttttgactaaaatctcaagactagcaagttctcttactaatttatatg
a M L L * K T D F R V L I V Q E N D * I Y -
b C C Y R K L I L E F * S F K R M I K Y T -
c V V I E N * F * S S D R S R E * L N I H -
atttcct
2581 ------- 2587
taaagga
a I S -
b F P -
c F -

# FIGURE 1



Enzymes that do cut:

Sau3AI

Enzymes that do not cut:

NONE

# Translate Tool - Results of translation

Please select one of the following frames:

5'3' Frame 1
X G R Q R A A P L S Q Y P D T L G T F P S H G F P G A D P L L E H N **Stop** H H H Y S G
W S N C T H H W L W Y F R E T L H H S H Y C R L S W E H W G G W N P E L H F **Stop** T
**Stop** H Q T F **Stop** Y R D T **Met** A E G R C F R L G P **Stop** V Q R R Q R **Stop** A V G A G
**Stop** N V Q R P D S S V C **Stop** S S D S W Q C L F A A E K R A T H R C W H L Q **Met** L Y
H H F **Stop** R Q G E C **Stop** P **Stop** V **Stop** N W S L Q H A G S E C G L **Stop** C Q L R D L A
V **Stop** G S P **Met** V P P A H S G L G I P S **Stop** P G S Q L L G S L Q Y Q L **Stop** A E L **Stop**
E C D H E G C V C A L Q C Y D Q Q H I L L Y D **Stop** K **Stop** H C Q S N R G Y Q S D R I
G D Q K A E S P T A A K L K G F S V C L F F L C H Q L G T S A S Q P L P D A K I **Met** C
L G H K K A C K V I V T T G I Y R T I S P P D **Met** T **Stop** F Y I S G R K **Stop** I H I **Stop** K
S G V S K Q E Q E T K R S Q K Q K A P I **Stop** T R **Stop** I Y L Q R H I R S W E N N S C E L
D K C V K S D K **Stop** N A R G D K C I P R S Q G P P P A C H L G S E R T G **Stop** C **Met** F
F V S E F L V I C A V **Met** L L **Stop** G S P W K V Y P N I S T S Y I P Q I K L **Stop** Y V P
**Stop** D A A N **Stop** L P L R N S G A A A F **Stop** **Stop** W V K **Stop** F T F Y D A S K G A L A
S L P N **Stop** Q **Met** P K L R K **Met** I I I L A **Stop** T E Q S A T P I L **Stop** I N **Stop** A P S F
**Stop** T N K C G F I S Q **Met** **Met** F I R E W S R E G P F T L T I W H Y V I T S S E A S P F H
P A W T A K T S V F N S I **Stop** S S G T Q L G **Stop** F R P P S P G E C L K T I L V T S **Met**
R E W R R I Q C Y Y Q L V D K G Q G C C S T S Y H V Q D V S P L Q L P N P K C Q L C
Q D **Stop** E T L V L S R K G P G K R G A N K S V C F S H **Stop** S L A N K H S V S L A A A
S A Q R A R T L S G T R I T S L S E Q S **Stop** Q G L W E **Met** P D G I I F S L L S F **Stop** V
S F P S F Y P A S Q V L **Stop** E K C L S S S S G F L T L N L D L Q T L P G H N S N **Stop** G
N K H I P S **Met** K H T Q T F E S K D N D C L N **Stop** G L E E **Stop** S F E G K E Y F V S S P
L P T L F **Met** C **Stop** P L P S W T L E P R **Stop** L Y Y **Met** L L **Stop** K T D F R V L I V Q E
N D **Stop** I Y I S

5'3' Frame 2
X E G S G Q L H S A S T Q I R W E P S P A **Met** A S L G Q I L F W S I I S I I I I L A G A I A
L I I G F G I S G R H S I T V T T V A S A G N I G E D G I L S C T F E P D I K L S D I V I Q
W L K E G V L G L V H E F K E G K D E L S E Q D E **Met** F R G R T A V F A D Q V I V G
N A S L R L K N V Q L T D A G T Y K C Y I I T S K G K G N A N L E Y K T G A F S **Met** P
E V N V D Y N A S S E T L R C E A P R W F P Q P T V V W A S Q V D Q G A N F S E V S
N T S F E L N S E N V T **Met** K V V S V L Y N V T I N N T Y S C **Met** I E N D I A K A T G
D I K V T E S E I K R R S H L Q L L N S K A S L C V S S F F A I S W A L L P L S P Y L
**Met** L K **Stop** C A L A T K K H A K S L L Q Q G S T E L F H H Q I **Stop** P S F I F L G G N E
F I S R S L E **Stop** A N K S K K Q K E A K S R R L Q Y E Q D K S I F K D I L E V G K I I H
V N **Stop** T S V L R V I S K **Met** H V E T S A S P D L R D L P L P V T W G V R G Q D S A
C S L S L N F **Stop** L Y V L **Stop** C C S E E A P G K S I P T Y P H L I F H K L S C S **Met** Y
P K T L L I D C H F A T Q G R L H F S N G S N D S L F **Met** **Met** L P K V P W L L F P T D

# FIGURE 2



K C Q S **Stop** E K **Stop** S **Stop** F **Stop** H K Q S S R R H R F Y K **Stop** T E H L L F K Q T N A
G L F L R **Stop** C S S V N G P G K D L S P **Stop** L Y G I **Met** S S Q A L R L L L S I L R G Q
L R P Q F S I A S R A V G L S W G D F A P H L R G N V **Stop** R Q F W L P Q **Stop** G S G G
G Y S A T T N **Stop** W I K A R D A A Q P P T **Met** Y R T S P H Y N Y P I R S V N C V R T
K K P W F **Stop** V E K G L E R G E P T N L S A S H I S H W Q I S I L S L W L L P Q H R E
P E L Y R A P G **Stop** H L S V N R V D K A Y G K C L **Met** G L S S A C **Stop** A S K F L S L
H S T L Q A K F C K R N A **Stop** V L A Q V F L L **Stop** I **Stop** I S R P F L A T I Q I K A T N
I Y L P **Stop** S T H R L L K A R T **Met** T A **Stop** I E A L R N E A L K E K N T L F P A P F P
H S S C V N H C L P G P W S H G D C I T C C Y R K L I L E F **Stop** S F K R **Met** I K Y T F
P

5'3' Frame 3

X K A A G S S T Q P V P R Y A G N L P Q P W L P W G R S S S G A **Stop** L A S S L F W L
E Q L H S S L A L V F Q Q G D T P S Q S L L S P Q L G T L G R **Met** E S **Stop** A A L L N L T
S N F L I S **Stop** Y N G **Stop** R K V F **Stop** A W S **Met** S S K K A K **Met** S C R S R **Met** K C
S E A G Q Q C L L I K **Stop** **Stop** L A **Met** P L C G **Stop** K T C N S Q **Met** L A P T N V I S S
L L K A R G **Met** L T L S I K L E P S A C R K **Stop** **Met** W T I **Met** P A Q R P C G V R L P D
G S P S P Q W S G H P K L T R E P T S R K S P I P A L S **Stop** T L R **Met** **Stop** P **Stop** R L C
L C S T **Met** L R S T T H T P V **Stop** L K **Met** T L P K Q Q G I S K **Stop** Q N R R S K G G V
T Y S C **Stop** T Q R L L C V S L L S L P S A G H F C L S A L T **Stop** C **Stop** N N V P W P Q
K S **Met** Q S H C Y N R D L Q N Y F T T R Y D L V L Y F W E E **Met** N S Y L E V W S E Q
T R A R N K K K P K A E G S N **Met** N K I N L S S K T Y **Stop** K L G K **Stop** F **Met** **Stop** T
R Q V C **Stop** E **Stop** **Stop** V K C T W R Q V H P Q I S G T S P C L S P G E **Stop** E D R I V
H V L C L **Stop** I F S Y **Met** C C N V A L R K P L E S L S Q H I H I L Y S T N **Stop** A V V C
T L R R C **Stop** L T A T S Q L R G G C I L V **Met** G Q **Met** I H F L **Stop** C F Q R C L G F S S
Q L T N A K V E K N D H N F S I N R A V G D T D F I N K L S T F F L N K Q **Met** R V Y F
S D D V H P **Stop** **Met** V Q G R T F H L D Y **Met** A L C H H K L **Stop** G F S F P S C V D S
**Stop** D L S F Q **Stop** H L E Q W D S A G V I S P P I S G G **Met** S E D N F G Y L N E G V E
E D T V L L P T S G **Stop** R P G **Met** L L N L L P C T G R L P I T T T Q S E V S T V S G L R
N P G F E **Stop** K R A W K E G S Q Q I C L L L T L V I G K **Stop** A F C L F G C C L S T E S
Q N S I G H Q D N I S Q **Stop** T E L T R P **Met** G N A **Stop** W D Y L Q L V E L L S F F P F I
L P C K P S S V R E **Met** P E F **Stop** L R F S Y S E F R S P D P S W P Q F K L R Q Q T Y T
F H E A H T D F **Stop** K Q G Q **Stop** L L E L R P **Stop** G **Met** K L **Stop** R K R I L C F Q P P S
H T L H V L T T A F L D L G A T V T V L H V V I E N **Stop** F **Stop** S S D R S R E **Stop** L N
I H F

3'5' Frame 1

R K C I F N H S L E R S E L **Stop** N Q F S I T T C N T V T V A P R S R K A V V N T **Stop** R
V W E G G W K Q S I L F L Q S F I P Q G L N S S S H C P C F Q K S V C A S W K V Y V C
C L N L N C G Q E G S G D L N S E **Stop** E N L S **Stop** N S G I S L T E L G L Q G R **Met** K
G K K L R S S T S **Stop** R **Stop** S H Q A F P I G L V N S V H **Stop** E **Met** L S W C P I E F W
L S V L R Q Q P K R Q N A Y L P **Met** T N V R S R Q I C W L P S F Q A L F Y S K P G F L
S P D T V D T S D W V V V **Met** G R R P V H G R R L S S I P G L Y P L V G S S T V S S S
T P S L R **Stop** P K L S S D I P P E **Met** G G E I T P A E S H C S R C Y **Stop** K L R S **Stop** L
S T Q D G K E K P Q S L **Stop** **Stop** H N A I **Stop** S R **Stop** K V L P W T I H G **Stop** T S S E
K **Stop** T R I C L F K K K V L S L F I K S V S P T A L F **Met** L K L **Stop** S F F S T L A F V S

# FIGURE 2



W E E K P R H L W K H H K K **Stop** I I **Stop** P I T K **Met** Q P P L S C E V A V N **Stop** Q R L
R V H T T A **Stop** F V E Y K **Met** W I C W D R L S R G F L R A T L Q H I **Stop** L K I Q R Q
R T C T I L S S H S P G D R Q G E V P E I W G C T C L H V H F T Y H S **Stop** H T C L V H
**Met** N Y F P N F **Stop** Y V F E D R F I L F I L E P S A F G F F L F L A L V C S L Q T S R Y
E F I S S Q K Y K T R S Y L V V K **Stop** F C R S L L **Stop** Q **Stop** L C **Met** L F C G Q G T L
F **Stop** H Q V R A E R Q K C P A D G K E R R D T Q R S L **Stop** V **Stop** Q L **Stop** V T P P F
D L R F C H F D I P C C F G N V I F N H T G V C V V D R N I V E H R H N L H G H I L R
V Q L K A G I G D F R E V G S L V N L G C P D H C G L G E P S G S L T P Q G L **Stop** A
G I I V H I H F R H A E G S S F I L K V S I P L A F R S D D I T F V G A S I C E L H V F Q P
Q R G I A N Y H L I S K H C C P A S E H F I L L R Q L I F A F F E L **Met** D Q A **Stop** N T F
L Q P L Y H D I R K F D V R F K S A A Q D S I L P N V P S **Stop** G D S S D C D G V S P
**Stop** N T K A N D E C N C S S Q N N D D A N Y A P E E D L P Q G S H G W G R F P A Y
L G T G **Stop** V E L P A A F X X

3'5' Frame 2

G N V Y L I I L L N D Q N S K I S F L **Stop** Q H V I Q S P W L Q G P G R Q W L T H E E C
G K G A G N K V F F S F K A S F L K A S I Q A V I V L A F K S L C V L H G R Y **Met** F V
A L I **Stop** I V A R K G L E I **Stop** I Q S K K T **Stop** A R T Q A F L L Q N L A C R V E **Stop**
R E R N L E A Q Q A E D N P I R H F P **Stop** A L S T L F T E R C Y P G A R **Stop** S S G S L
C **Stop** G S S Q R D R **Met** L I C Q **Stop** L **Met** **Stop** E A D R F V G S P L S R P F S T Q N Q
G F L V L T Q L T L R I G **Stop** L **Stop** W G D V L Y **Met** V G G **Stop** A A S L A F I H **Stop**
L V V A L Y P P P L P H **Stop** G N Q N C L Q T F P R R W G A K S P Q L S P T A L D A I E
N **Stop** G L S C P R R **Met** E R R S L R A C D D I **Met** P Y S Q G E R S F P G P F T D E H H
L R N K P A F V C L K R R C S V Y L **Stop** N R C R R L L C L C **Stop** N Y D H F S Q L W
H L S V G K R S Q G T F G S I I K S E S F D P L L K C S R P **Stop** V A K W Q S I S S V L G
Y I L Q L N L W N I R C G Y V G I D F P G A S S E Q H Y S T Y N **Stop** K F R D K E H A L
S C P L T P Q V T G R G R S L R S G D A L V S T C I L L I T L N T L V **Stop** F T **Stop** I I F
P T S N **Met** S L K I D L S C S Y W S L L L L A S F C F L L L F A H S R L L D **Met** N S F P
P R N I K L G H I W W **Stop** N S S V D P C C N N D F A C F F V A K A H Y F S I R **Stop** G
L R G R S A Q L **Met** A K K E E T H R E A F E F S S C R **Stop** L R L L I S D S V T L I S P V
A L A **Met** S F S I I Q E Y V L L I V T L **Stop** S T D T T F **Met** V T F S E F S S K L V L E T
S E K L A P W S T W D A Q T T V G W G N H R G A S H R K V S E L A L **Stop** S T F T S G
**Met** L K A P V L Y S R L A F P L P L E V **Met** I **Stop** H L **Stop** V P A S V S C T F F S R K E
A L P T I T **Stop** S A N T A V R P L N I S S C S D S S S L P S L N S W T K P K T P S F S H
C I T I S E S L **Met** S G S K V Q L R I P S S P **Met** F P A E A T V V T V **Met** E C L P E I P K
P **Met** **Met** S A I A P A R I **Met** **Met** **Met** L I **Met** L Q K R I C P R E A **Met** A G E G S Q R I
W V L A E W S C P L P S X

3'5' Frame 3

E **Met** Y I **Stop** S F S **Stop** T I R T L K S V F Y N N **Met** **Stop** Y S H R G S K V Q E G S G
**Stop** H **Met** K S V G R G L E T K Y S F P S K L H S S R P Q F K Q S L S L L S K V C V C F
**Met** E G I C L L P **Stop** F E L W P G R V W R S K F R V R K P E L E L R H F S Y R T W L
A G **Stop** N E G K E T **Stop** K L N K L K I I P S G I S H R P C Q L C S L R D V I L V P D R
V L A L C A E A A A K E T E C L F A N D **Stop** C E K Q T D L L A P L F P G P F L L K T R
V S **Stop** S **Stop** H S **Stop** H F G L G S C N G E T S C T W **Stop** E V E Q H P W P L S T S W
**Stop** **Stop** H C I L L H S L I E V T K I V F R H S P G D G G R N H P S **Stop** V P L L **Stop**

# FIGURE 2



**Met** L L K T E V L A V H A G W K G E A S E L V **Met** T **Stop** C H I V K V K G P S L D H S
R **Met** N I I **Stop** E I N P H L F V **Stop** K E G A Q F I Y K I G V A D C S V Y A K I **Met** I I F
L N F G I C Q L G R E A K A P L E A S **Stop** K V N H L T H Y **Stop** N A A A P E L R S G S
Q L A A S **Stop** G T Y Y S L I C G I **Stop** D V D **Met** L G **Stop** T F Q G L P Q S N I T A H I T
K N S E T K N **Met** H Y P V L S L P R **Stop** Q A G G G P **Stop** D L G **Met** H L S P R A F Y L
S L L T H L S S S H E L F S Q L L I C L **Stop** R **Stop** I Y L V H I G A F C F W L L F V S C S
C L L T P D F **Stop** I **Stop** I H F L P E I **Stop** N **Stop** V I S G G E I V L **Stop** I P V V T **Met** T
L H A F L W P R H I I L A S G K G **Stop** E A E V P S **Stop** W Q R K K R H T E K P L S L A
A V G D S A F **Stop** S P I L S L **Stop** Y P L L L W Q C H F Q S Y R S **Met** C C **Stop** S **Stop**
H C R A Q T Q P S W S H S Q S S A Q S W Y W R L P R S W L P G Q L G **Met** P R P L W A
G G T I G E P H T A R S L S W H Y S P H S L P A C **Stop** R L Q F Y T Q G **Stop** H S P C L
**Stop** K **Stop** **Stop** Y N I C R C Q H L **Stop** V A R F S A A K R H C Q L S L D Q Q T L L S G
L **Stop** T F H P A P T A H L C L L **Stop** T H G P S L K H L P S A I V S R Y Q K V **Stop** C Q
V Q K C S S G F H P P Q C S Q L R R Q **Stop** **Stop** L **Stop** W S V S L K Y Q S Q **Stop** **Stop**
V Q L L Q P E **Stop** **Stop** **Stop** C **Stop** L C S R R G S A P G K P W L G K V P S V S G Y W L
S G A A R C L X X

# FIGURE 3

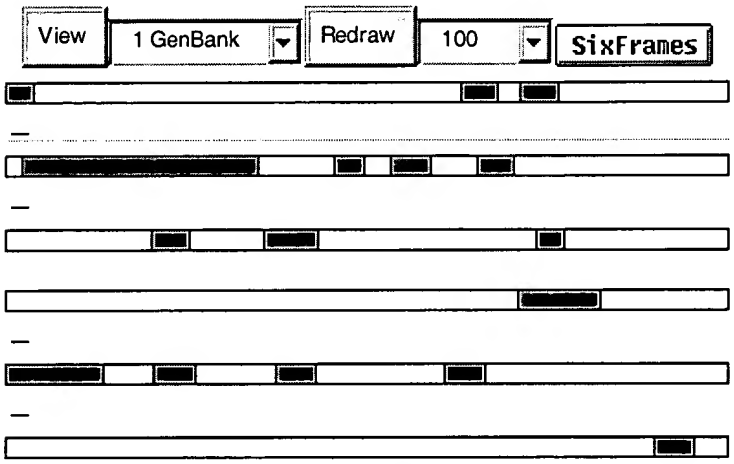


NCBI **ORF Finder (Open Reading Frame Finder)**

PubMed Entrez BLAST OMIM Taxonomy Structure

## DEX0043_1

Program blastp Database nr BLAST with parameters Cognitor




Length: **282 aa**

Accept | Alternative Initiation Codons

| Frame | from | to | Length |
|---|---|---|---|
| +2 | 62 | 910 | 849 |
| -2 | 1 | 354 | 354 |
| -1 | 1835 | 2134 | 300 |
| +3 | 933 | 1127 | 195 |
| -2 | 1576 | 1725 | 150 |
| -2 | 973 | 1122 | 150 |
| -2 | 535 | 684 | 150 |
| -3 | 2328 | 2471 | 144 |
| +2 | 1382 | 1525 | 144 |
| +1 | 1843 | 1980 | 138 |
| +3 | 528 | 665 | 138 |
| +1 | 1633 | 1767 | 135 |
| +2 | 1691 | 1822 | 132 |
| +2 | 1184 | 1291 | 108 |
| +3 | 1899 | 2000 | 102 |
| +1 | 1 | 102 | 102 |

```
 62 atggcttccctggggcagatcctcttctggagcataattagcatc
    M  A  S  L  G  Q  I  L  F  W  S  I  I  S  I
107 atcattattctggctggagcaattgcactcatcattggctttggt
    I  I  I  L  A  G  A  I  A  L  I  I  G  F  G
152 atttcagggagacactccatcacagtcactactgtcgcctcagct
    I  S  G  R  H  S  I  T  V  T  T  V  A  S  A
197 gggaacattggggaggatggaatcctgagctgcacttttgaacct
    G  N  I  G  E  D  G  I  L  S  C  T  F  E  P
242 gacatcaaactttctgatatcgtgatacaatggctgaaggaaggt
    D  I  K  L  S  D  I  V  I  Q  W  L  K  E  G
287 gttttaggcttggtccatgagttcaaagaaggcaaagatgagctg
    V  L  G  L  V  H  E  F  K  E  G  K  D  E  L
332 tcggagcaggatgaaatgttcagaggccggacagcagtgtttgct
    S  E  Q  D  E  M  F  R  G  R  T  A  V  F  A
377 gatcaagtgatagttggcaatgcctctttgcggctgaaaaacgtg
    D  Q  V  I  V  G  N  A  S  L  R  L  K  N  V
422 caactcacagatgctggcacctacaaatgttatatcatcacttct
    Q  L  T  D  A  G  T  Y  K  C  Y  I  I  T  S
467 aaaggcaaggggaatgctaaccttgagtataaaactggagccttc
    K  G  K  G  N  A  N  L  E  Y  K  T  G  A  F
512 agcatgccggaagtgaatgtggactataatgccagctcagagacc
    S  M  P  E  V  N  V  D  Y  N  A  S  S  E  T
```

# FIGURE 3



```
557 ttgcggtgtgaggctccccgatggttcccccagcccacagtggtc
    L  R  C  E  A  P  R  W  F  P  Q  P  T  V  V
602 tgggcatcccaagttgaccagggagccaacttctcggaagtctcc
    W  A  S  Q  V  D  Q  G  A  N  F  S  E  V  S
647 aataccagctttgagctgaactctgagaatgtgaccatgaaggtt
    N  T  S  F  E  L  N  S  E  N  V  T  M  K  V
692 gtgtctgtgctctacaatgttacgatcaacaacacatactcctgt
    V  S  V  L  Y  N  V  T  I  N  N  T  Y  S  C
737 atgattgaaaatgacattgccaaagcaacaggggatatcaaagtg
    M  I  E  N  D  I  A  K  A  T  G  D  I  K  V
782 acagaatcggagatcaaaaggcggagtcacctacagctgctaaac
    T  E  S  E  I  K  R  R  S  H  L  Q  L  L  N
827 tcaaaggcttctctgtgtgtctcttctttctttgccatcagctgg
    S  K  A  S  L  C  V  S  S  F  F  A  I  S  W
872 gcacttctgcctctcagcccttacctgatgctaaaataa 910
    A  L  L  P  L  S  P  Y  L  M  L  K  *
```

# What's New Archive



## 1996

***What's New Archive***

2001
2000
1999
1998
1997
1996
1995
1994

BACK

| | | |
|---|---|---|
| 11/25 | Electronic PCR | Electronic PCR is now available. PCR-based sequence tagged sites (STSs) have been used as landmarks for construction of various types of genomic maps. Using "electronic PCR" (e-PCR), these sites can be detected in DNA sequences, potentially allowing their map locations to be determined. |
| 11/14 | Sequin, Release 1.71 | A new release of Sequin, a sequence submission tool, is now available. Version 1.71 features improved handling of phylogenetic sets of sequences and also allows users to update their own pre-existing database records. |
| 11/04 | dbGSS Announced | The Database of Genome Survey Sequences (dbGSS) is now available. This database contains more detailed information than the corresponding records in the GSS Division of GenBank. |
| 10/24 | Human Gene Map | The Gene Map of the Human Genome published in the October 25 issue of *Science* is available. This map shows the chromosome location of over 16,000 human genes with links to the underlying sequence and map data. |
| 10/04 | Sequin | Sequin, a stand-alone sequence submission tool, has a new release with several enhancements, including a repeat finder and ORF finder. New documentation and a tutorial are available, both on the Web and in NCBI's newsletter. |
| 09/27 | ORF Finder | The Open Reading Frame (ORF) Finder is a graphical analysis tool that finds all open reading frames in a user's sequence or one already in the database of a selectable minimum size. |
| 09/06 | Virological Software | Software for analyzing animal trials and calculating infectious and 50% inhibitory doses is now available. The programs VacMan and |

| | | |
|---|---|---|
| | | ID-50 can now be downloaded as self-extracting archives for either IBM or Macintosh computers. |
| 08/23 | Complete Genome, *Methanococcus jannaschii* | The complete genome sequence and annotation of *Methanococcus jannaschii*, prepared by The Institute for Genomic Research (TIGR) is now available in Entrez Genome, as well as in GenBank, where the 1.7-megabase sequence has been separated into 150 records of approximately 11,000 bp each. The graphical view (as well as a link to underlying data) of the complete genome is present in Entrez Genome, along with the extrachromosomal elements 1 and 2. The complete sequence is also available by anonymous FTP; see the README file for a description of the various files in the genomes division directory. |
| 08/20 | Batch Entrez | Downloading large numbers of sequence records from Entrez is now possible through 'Batch Entrez'. User can specify a download for an entire set of records for a given organism or for a set of accession numbers. The data are saved to a file on the user's computer. |
| 08/05 | *Saccharomyces cerevisiae* Database | A new database has been added to the BLAST databases: all the nucleotide sequences from the yeast (*Saccharomyces cerevisiae*) genome sequencing project and their encoding amino acid sequences can now be searched with the BLAST suite of programs. |
| 07/26 | Cn3D in Entrez | A major new release of Network Entrez is now available. Release 5.0 contains Cn3D, a new 3D structure viewer integrated into Network Entrez. |
| 07/15 | BLAST2 | The BLAST2 network service is now available on the FTP site without registration. Three clients for multiple platforms are available: blastcli has a convenient graphical interface and produces the "traditional" BLAST output; blastcl2 is a command-line client (meant mostly for UNIX) that also produces the traditional BLAST output; and PowerBlast produces a one-to-many alignment, allows filtering by organism, and allows a gapped alignment as a post-processing of the BLAST results. Users of the older Experimental BLAST Network Service (with the exception of GCG users, who are still required to register |

| | | |
|---|---|---|
| | | and use the older program) are encouraged to switch to this newest version. |
| 05/21; see also 11/14 | Sequin | Sequin is a program for submitting and updating GenBank entries. It is designed to simplify the sequence submission process, provide graphical viewing and editing options, and allow submission of segmented entries. Sequin automatically adjusts feature table positions as the sequence is edited. Versions of Sequin are available through FTP for the Macintosh, PC/Windows, UNIX, and VMS. |
| 05/21 | PowerBlast | PowerBlast is a new network BLAST application for automated analysis of genomic sequences. It combines BLAST searching with filtering for low complexity regions and repeats. It can generate organism-specific output and compute optimal, gapped alignments. The results are displayed graphically and textually as multiple alignments, with annotated features superimposed on the aligned sequences. Versions of PowerBlast are available through FTP for the Macintosh, PC, SunOS, and Solaris. |
| 05/06 | WWW BLAST | The WWW BLAST page has been extensively revised. It now has both a simplified "Basic" Blast Search, allowing a user to search with the default parameters, as well as an "Advanced" page, where users may set BLAST parameters. An email option allows a user to receive results in a convenient form. |
| 04/10 | WWW Entrez | WWW Entrez now provides graphical views of nucleotide and protein sequences and access to the NCBI Genomes database, which contains graphical views of sequences and chromosome maps. Click on "Graphical view" from an Entrez document summary or click on the "Graphic" button from a sequence report. |
| 03/12 | Mouse/Human Homology | The Seldin/Debry Mouse/Human Homology Relationships page presents a table comparing genes in homologous segments of DNA from human and mouse sources, sorted by position in each genome. |
| 03/07 | Complete Genomes | An NCBI research project, Complete Genomes, presents the results of analyses of complete genome sequences. The analyses for the genomes of *Haemophilus influenzae*, *E. coli* (75%), and *Mycoplasma genitalium* are |

| | | |
|---|---|---|
| | | now available. |
| 03/08 | BLAST Databases | Changes to the BLAST Databases (February 20 announcement superseded by that of March 8.) |
| 02/15 | Homepage Reorganization | Major reorganization of the NCBI homepage with new top-level links to additional databases and services. |
| 02/07 | International Database Collaboration | The International Nucleotide Sequence Database Collaboration page describes current projects and provides links to the sites. |
| 01/30 | NCBI Structure Group | The NCBI Structure Group (Steve Bryant) has a new page providing access to their structure research, the PKB and MMDB databases, and threading software. |

Revised: June 6, 2002.

Volume 12 Number 1 1984 Nucleic Acids Research

# A comprehensive set of sequence analysis programs for the VAX

John Devereux, Paul Haeberli* and Oliver Smithies

Laboratory of Genetics, University of Wisconsin, Madison, WI 53706, USA

Received 18 August 1983

ABSTRACT

The University of Wisconsin Genetics Computer Group (UWGCG) has been organized to develop computational tools for the analysis and publication of biological sequence data. A group of programs that will interact with each other has been developed for the Digital Equipment Corporation VAX computer using the VMS operating system. The programs available and the conditions for transfer are described.

## INTRODUCTION

The rapid advances in the field of molecular genetics and DNA sequencing have made it imperative for many laboratories to use computers to analyze and manage sequence data. UWGCG was founded when it became clear to several faculty members at the University of Wisconsin that the there was no set of sequence analysis programs that could be used together as a coherent system and be modified easily in response to new ideas.

With intramural support a computer group was organized to build a strong foundation of software upon which future programs in molecular genetics could be based. This initial project has been completed and the resulting programs, written in Fortran 77, are available for VAX computers using the VMS operating system. Most of the programs can be used with only a terminal, although several require a Hewlett Packard plotter.

UWGCG software has been installed for testing at eight different institutions. A simple method has been developed for transferring and maintaining this system on other VAX computers.

## DESIGN PRINCIPLES

UWGCG program design is based on the "software tools" approach of Kernighan and Plauger(1). Each program performs a simple function and is easy to use. The programs can be used independently in different combinations so

that complex problems are solved by the use of several programs in succession. New programming is simplified since less effort is required to bridge a gap between existing programs.

UWGCG software is designed to be maintained and modified at sites other than the University of Wisconsin. The program manual is extensive and the source codes are organized to make modification convenient. Scientists using UWGCG software are encouraged to use existing programs as a framework for developing new ones. Our copyright can be removed from any program modified by more than 25% of our original effort.

## PROGRAMS AVAILABLE FROM UWGCG

The programs described below are named and defined individually in Table 1. Program names in the text are underlined.

### Comparisons

Comparisons may be done with "dot plots" using the method of Maizel and Lenk(2). Optimal alignments can be generated by the methods of Needleman and Wunsch(3), of Sellers(4), and the "local homology" method of Smith and Waterman(5). The Smith and Waterman alignment algorithm is also the most sensitive method available for identifying similarities between weakly related sequences.

### Mapping and Searching

Mapping is available in several formats. Graphic maps display all of the cuts for each restriction enzyme on parallel lines. This graphic map facilitates selection of enzymes for isolating any region of a sequenced DNA molecule. Sorted maps in tabular format arrange the fragments from any digestion in order of molecular weight to show which fragments are similar in size and thus likely to be confused in gels. Another frequently used mapping format, designed by Frederick Blattner(6), displays the enzyme cuts above the original DNA sequence. Both strands of the DNA and all six frames of translation are shown.

All mapping programs will search for user-specified sequences, allowing features to be marked at the appropriate position on a restriction map. The mapping and searching programs can be used to aid site-specific mutagenesis experiments by showing where mutations could generate new restriction sites. All of the positions in a sequence where a synthetic probe could pair with one or more mismatches can also be located. Sequences related to less precisely defined features such as promoters or intervening sequence splice sites, can be located with a program that uses a consensus sequence as a probe. The

Table 1

Programs Available from UWGCG

| Name | Function |
|---|---|
| DotPlot+ | makes a dot plot by method of Maizel and Lenk(2) |
| Gap | finds optimal alignment by method of Needleman and Wunsch(3) |
| BestFit | finds optimal alignment by method of Smith and Waterman(5) |
| MapPlot+ | shows restriction map for each enzyme graphically |
| MapSort | tabulates maps sorted by fragment position and size |
| Map | displays restriction sites and protein translations above and below the original sequence(Blattner,6) |
| Consensus | creates a consensus table from pre-aligned sequences |
| FitConsensus | finds sequences similar to a consensus sequence using a consensus table as a probe |
| Find | finds sites specified interactively |
| Stemloop | finds all possible stems (inverted repeats) and loops |
| Fold* | finds an RNA secondary structure of minimum free energy by the method of Zuker(7) |
| CodonPreference+ | plots the similarity between the codon choices in each reading frame and a codon frequency table(8) |
| CodonFrequency | tabulates codon frequencies |
| Correspond | finds similar patterns of codon choice by comparing codon frequency tables (Grantham et al,9) |
| TestCode+ | finds possible coding regions by plotting the "TestCode" statistic of Fickett(10) |
| Frame+ | plots rare codons and open reading frames(8) |
| PlotStatistics+ | plots asymmetries of composition for one strand |
| Composition | measures composition, di and trinucleotide frequencies |
| Repeat | finds repeats (direct, not inverted) |
| Fingerprint | shows the labelled fragments expected for an RNA fingerprint |
| Seqed | screen oriented sequence editor for entering, editing and checking sequences |
| Assemble | joins sequences together |
| Shuffle | randomizes a sequence maintaining composition |
| Reverse | reverses and/or complements a sequence |
| Reformat | converts a sequence file from one format to another |
| Translate | translates a nucleotide into a peptide sequence |
| BackTranslate | translates a peptide into a nucleotide sequence |
| Spew | sends a sequence to another computer |
| GetSeq | accepts a sequence from another computer |
| Crypt | encrypts a file for access only by password |
| Simplify | substitutes one of six chemically similar amino acid families for each residue in a peptide sequence |
| Publish | arranges sequences for publication |
| Poster+ | plots text (for labelling figures and posters) |
| OverPrint | prints darkened text for figures with a daisy wheel printer |

+ requires a Hewlett Packard Series 7221 terminal plotter
* Fold is distributed by Dr. Michael Zuker not UWGCG.

mapping programs can also be used on protein sequences to identify the peptides resulting from proteolytic cleavage.

Secondary Structure

Three programs are available to examine secondary structure in nucleic acids. The program StemLoop identifies all inverted repeats. An implementation of Dr. Michael Zuker's Fold program(7) finds an RNA secondary structure of minimum free energy based on published values of stacking and loop destabilizing energies. The "dot plot" comparison (mentioned above) of a sequence compared to its opposite strand gives a graphic picture of the pattern of inverted repeats in a sequence.

Analysis of Composition and the Location of Genetic Domains

Regions of a sequence with non-random base distribution can be displayed with three graphic tools designed to identify genetic domains. The program CodonPreference(8) identifies potential coding regions by searching through each reading frame for a pattern of preferred codon choices. The CodonPreference plot predicts the level of translational expression of mRNAs and helps identify frame shifts in DNA sequence data. Patterns of codon choice can be compared with the program Correspond(9). When a strong pattern of codon preferences is not expected, the "TestCode" statistic of Fickett(10) can be plotted to show regions of compositional constraint at every third base. Another program plots asymmetries of composition by strand. Strand asymmetries have been associated with genetic domains by several authors(11)(12). A fourth program called Frame marks the positions of rare codons and open reading frames on a graph showing all six reading frames.

Several tools are available to measure content and to count dinucleotide, trinucleotide, neighbor and repeat frequencies. A program that predicts RNA fingerprint patterns and another that tabulates codon frequencies complete the group of programs that analyze composition.

Sequence Manipulation

Sequences may be entered, assembled, edited, reversed, randomized, reformatted, translated, back-translated, documented, transferred, or encrypted rapidly with a large set of sequence manipulation tools.

A screen-oriented editor is available that allows sequences to be entered and checked. After a sequence is entered, it may be reentered for proofreading. Whenever a reentered base is at variance with the original, the terminal bell rings and the position is marked. Existing sequences can be edited quickly by moving directly to a sequence position specified by either a coordinate or a sequence pattern. The program can reassign the terminal's

keys to place G, A, T and C conveniently under the fingers of one hand in the same order as the lanes of a sequencing gel.

Programs are available for changing sequence file format. Sequence data from any source can be used in UWGCG programs, and sequence files maintained with UWGCG software can be converted for use in other non-UWGCG programs. For instance, the programs of Roger Staden(13) or Intelligenetics Inc.(14) could be used to assemble a sequence from the sequences of many small sub-fragments generated by DNAase I digestion. The assembled sequence could then be reformatted for use in any UWGCG program. A program is available that transfers sequences to and from other computers.

Sequence Publication

A program, Publish, will format sequences into figures. Publish has alternatives for line size, numbering, scaling, translation and comparison to other sequences. Poster is a program that will plot text on figures.

## GENERAL FEATURES OF UWGCG SOFTWARE

### Interactive Style

Each program is run by simply typing its name. Every parameter required by the program is obtained interactively. Questions are answered with a file name, a yes, a no, a number, or a letter from a menu. Default answers are displayed. Programs are insensitive to absurd answers and will ask the question again if, for instance, you name a file that does not exist or if you use a nonnumeric character when typing a number. Special features such as plotting features oriented to publication, are obtained by using an extra word next to the program's name when the program is run. Thus parameter queries are kept to a minimum for the normal use of each program.

### Data

Both the NIH-GenBank(15) and the EMBL(16) nucleotide sequence data libraries are available "on-line" to any UWGCG program. A Search utility will locate sequences in the libraries by key word. A Find utility will locate library entries containing any specified sequence. A program is available that installs the new data sent periodically from GenBank and EMBL to update their data libraries.

All of the data in the system are stored in text files that can be read and modified easily. Every data file has an English heading describing the contents. The data files may be copied by each user for analysis or modification. Programs recognize and read user-modified input data automatically. Data files can be modified with any text editor.

Sequence File Structure

Sequences are maintained in files that allow documentation and numbering both above and within the sequence. This file format is compatible with both of the nucleic acid sequence libraries and has been adopted as the standard sequence file format by the data base project at the European Molecular Biology Lab. Because genetic manipulations commonly involve linking several molecules of known sequence, UWGCG sequence files are designed to support concatenation by allowing comments to appear within the sequences at any location. Coding sequences or the boundaries between cloning vector and insert, for instance, can be marked within the sequence itself for immediate identification.

Sequence Symbols

All possible nucleotide ambiguities and all standard one-letter amino acid codes are part of the UWGCG symbol set that includes all alphabetic characters plus five additional characters. The proposed IUB-IUPAC standard nucleotide ambiguity symbols(17) are used for the mapping, searching and comparison programs. Lower case characters are used in sequences to indicate uncertainty as distinct from ambiguity. This allows the entire lexicon of symbols to be reused with same meaning, but with the prefix "maybe-." This reuse of the symbol set in lower case makes the uncertainty symbols more complete, understandable and visible.

Symbol Comparison

Sequence analysis programs generally make comparisons between sequence symbols (bases or amino acids) in order to find enzyme sites, create alignments, locate inverted repeats etc. These symbol comparisons are handled in several ways.

Symbol comparisons for alignment, comparison and secondary structure analysis are made by looking up a value in a symbol comparison table for the quality of the match. The table might contain 1's for matches and 0's for mismatches. If amino acids are being compared, however, a real number could be assigned at each position based on some previously assigned chemical similarity of the pair of residues or on the mutational distance between their codons. Standard symbol tables are provided by UWGCG, but the system is designed to allow each user to specify his own values.

Symbols comparisons for mapping and searching operations in nucleic acids are made by converting the IUB-IUPAC symbols into a binary code. The bits of this code represent G, A, T and C with ambiguity symbols causing more than one

bit to be set. A group of library functions identify overlap between the bits for each IUB-IUPAC symbol.

Documentation

Documentation is available both in printed form and on the terminal screen. A 350 page manual describes the operation of each program in detail, gives practical considerations and shows what will appear on the screen during a session with the program. Output files and plots are shown for the session. The data for the session shown in the documentation are included with the system so that the each program's operation can be checked. The "on-line" documentation is the same as the manual, but can be changed immediately when a program is modified.

All programs write output to files that are completely documented and sensibly organized for input to other programs. The input data, the program and the parameters used are clearly identified in every output file.

Procedure Library

UWGCG programs are written largely as calls to a library of 250 procedures designed to manipulate biological sequences. These procedures use data and file structures which have been designed to simplify program modification. For instance, standard operations such as reading sequences from files are always handled by a single library procedure. Thus a change in sequence file format requires only one subroutine to be modified for the new format to be acceptable to all of the programs in the system. Command procedures are available to help modify the library. The procedure library can be used by programs written in any language.

## DISTRIBUTION OF UWGCG SOFTWARE

Intent

The intent of UWGCG is to make its software available at the lowest possible cost to as many scientists as possible.

Fees

A fee of $2,000 for non-profit institutions or $4,000 for industries is being charged for a tape and documentation for each computer on which UWGCG software is installed. While no continuing fee is required, UWGCG software, like the field it supports, is changing very rapidly. A consortium of industries and academic laboratories is planned to support the project in the future. The consortium will entitle its members to periodic updates and to influence the direction of new programming undertaken by UWGCG in return for a pledge of continuing financial support.

Copyrights

UWGCG retains the copyrights to all of its software and UWGCG must be contacted before all or any part of the its software package is copied or transferred to any machine. UWGCG is, however, mandated to provide research tools to help scientists working in the area of molecular genetics and we are glad to see our source codes become the basis of further programming efforts by other scientists. Copyright can be removed for any program modified by more than 25% of its original effort.

Tape Format

The UWGCG package is usually distributed in VAX/VMS "backup" format on a 9 track magnetic tape recorded at 1600 bits/inch. The system consists of about 1000 files using about 20,000 blocks at 512 bytes/block. The current versions of the GenBank and EMBL nucleotide sequence data bases are normally included which add another 3,000 files and require another 20,000 blocks.

Upon request UWGCG will make a card image tape of all of the Fortran 77 programs and procedures for reading on computers other than the VAX. The card image tape is usually provided at 1600 bits/inch with 80 characters/record and 10 records/block. Adaptation of UWGCG software to systems other than VAX/VMS may take considerable effort.

Equipment Required

UWGCG programs and command procedures will run on a Digital Equipment Corporation (DEC) VAX computer that is using version 3.0 or greater of the DEC VMS operating system. A tape drive is necessary; a floating point accelerator and a DEC Fortran compiler are helpful, but not required. All programs can be run from a DEC VT52 or VT100 terminal. Seven programs, as noted in table 1, require a Hewlett Packard 7221 terminal plotter wired in series with the terminal. Several utilities support a daisy wheel compatible printer attached to the terminal's pass-through port, however, all programs write output files suitable for printing on any standard device.

Inquiries

Inquiries may be sent to John Devereux at the Laboratory of Genetics, University of Wisconsin, Madison, WI, USA 53706, (608) 263-8970. UWGCG is not licensed to distribute Fold(7), but the UWGCG implementation is available from Michael Zuker, Division of Biological Sciences, National Research Council of Canada, 100 Sussex Drive, Ottawa, Canada, K1A OR6 (613) 992-4182.

ACKNOWLEDGEMENTS

UWGCG was started with software written for Oliver Smithies' laboratory

with NIH support from grants GM 20069 and AM 20120. UWGCG is directed by John Devereux and is operated as a part of the Laboratory of Genetics with the advice of a steering committee consisting of Richard Burgess, James Dahlberg, Walter Fitch, Oliver Smithies and Millard Susman. UWGCG is currently supported with intramural funds and with fees paid by the faculty and industries using the facility in Madison. This article is paper number 2684 from the Laboratory of Genetics, University of Wisconsin.

*Current address: Silicon Graphics Inc., 630 Clyde Court, Mountain View, CA 94043, USA

REFERENCES

1. Kernighan, B.W. and Plauger, P.J. (1976) Software Tools, Addison-Wesley Publishing Company, Reading, Massachusetts.
2. Maizel, J.V. and Lenk, R.P. (1981) Proceedings of the National Academy of Sciences USA 78, 7665-7669.
3. Needleman, S.B. and Wunsch, C.D. (1970) Journal of Molecular Biology 48, 443-453.
4. Sellers, P.H. (1974) SIAM Journal on Applied Mathematics 26, 787-793.
5. Smith, T.F. and Waterman, M.S. (1981) Advances in Applied Mathematics 2, 482-489.
6. Schroeder, J.L. and Blattner, F.R. (1982) Nucleic Acids Research 10, 69-84, Figure 1.
7. Zuker, M. and Stiegler, P. (1981) Nucleic Acids Research 9, 133-148.
8. Gribskov, M., Devereux, J. and Burgess, R.R. "The Codon Preference Plot: Graphic Analysis of Protein Coding Sequences and Gene Expression," submitted to Nucleic Acids Research.
9. Grantham, R. Gautier, C. Guoy, M. Jacobzone, M. and Mercier R. (1981) Nucleic Acids Research 9(1), r43-r74.
10. Fickett, J.W. (1982) Nucleic Acids Research 10, 5303-5318
11. Smithies, O., Engels, W.R., Devereux, J.R., Slightom, J.L., and S. Shen, (1981) Cell 26, 345-353.
12. Smith, T.F., Waterman, M.S. and Sadler, J.R. (1983) Nucleic Acids Research 11, 2205-2220.
13. Staden, R. (1980) Nucleic Acids Research 8, 3673-3694.
14. Clayton, J. and Kedes, L. (1982) Nucleic Acids Research 10, 305-321.
15. The GenBank(TM) Genetic Sequence Data Bank is available from Wayne Rindone, Bolt Beranek and Newman Inc., 10 Moulton Street, Cambridge, Massachusetts 02238, USA.
16. The EMBL Nucleotide Sequence Data Library is available from Greg Hamm, European Molecular Biology Laboratory, Postfach 10.2209, Meyerhofstrasse 1, 6900 Heidelberg, West Germany.
17. Personal communication from Dr. Richard Lathe, Transgene SA, 11 Rue Humann, 67000 Strasbourg, France.

# The ExPASy Molecular Biology Server

## History of changes, improvements and new features

If you subscribe to our Swiss-Flash service of electronic bulletins, you can receive these and other news by electronic mail.

October 14, 2004

- *Tools*
  Aldente is a tool to identify proteins from peptide mass fingerprinting data. This new, fast and powerful PMF tool uses the Hough transform to determine the mass spectrometer deviation, to realign the experimental masses and to exclude outliers ( More information).

- *Mirror site*
  We are happy to announce a new **ExPASy mirror site in Brazil**, http://br.expasy.org/, hosted by the Laboratório Nacional de Computação Científica, Petropolis

June 4, 2004

- The Melanie page has been restyled. It has been redesigned by the occasion of the announcement of SIB, Genebio and Amersham Biosciences joining forces to create one single 2D image analysis. Melanie was chosen to be integrated into ImageMaster™ 2D Platinum software.

April 14, 2004

- *UniProt*
  Since the last Swiss-Flash Bulletin, the universal protein resource, UniProt has been released publically. Many ExPASy pages and services have changed to accommodate different aspects of the UniProt knowledgebase and UniRef, the non-redundant reference databases of UniProt.

  In particular, the ExPASy BLAST interface now allows to launch a sequence similarity search against the UniRef clusters UniRef100, UniRef90 and UniRef50.

  Implicit links to UniRef50 and UniRef90 have been added to the NiceProt view of UniProt knowledgebase entries.

- *FTP server structure*
  As announced in a previous Swiss-Flash bulletin, the structure of the ExPASy ftp server has changed. Please refer to this previous announcement for details.

- *Swiss-Prot/TrEMBL (UniProt knowledgebase)*
  A note to Swiss-Prot and TrEMBL users: Please note that we have a long list of planned format changes to be introduced in the next few months.

  In the NiceProt view for Swiss-Prot/TrEMBL entries we have added implicit links to the Swiss-Model repository of 3D homology models (SMR).

  It is now possible to submit all splice isoforms annotated in one Swiss-Prot entry to a multiple

alignment, or to retrieve the sequences of all these isoforms, e.g. from http://www.expasy.org/cgi-bin/niceprot.pl?P29590#comments or from http://www.expasy.org/cgi-bin/get-varsplic.pl?P29590-4

- *PROSITE*
  PSview The view of PROSITE documentation entries contains new functionalities. When a 3D structure is described in the text, a direct link to a 3D image of the domain is provided. The Swiss-Prot match list of each signature can be visualized as a multiple alignment, or as a taxonomic distribution graph. For PROSITE profiles, a domain arrangement view is also provided where active sites and disulfide bridges annotated in Swiss-Prot entries are superimposed on PROSITE domains. see the following links for more details: http://www.expasy.org/cgi-bin/nicedoc.pl?PDOC50119 http://www.expasy.org/cgi-bin/prosite/PSView.cgi?ac=PS50119&onebyarch=1&trembl=1&hscale=0.6

- *ENZYME*
  Access to ENZYME entries by class, subclass etc. has been improved. It is now possible to easily retrieve all ENZYME or Swiss-Prot entries corresponding to a given ENZYME class. This functionality is available from a given ENZYME entry or for a given ENZYME class.

  The legends for the Biochemical Pathways have been made available in html and pdf format.

- *Tools*
  Myristoylator is a new tool designed to predict N-terminal myristoylation of proteins by neural networks. N-terminal myristoylation is a post-translational modification that causes the addition of a myristate group to the N-terminal glycine of an amino acid chain.

September 26, 2003

- *Swiss-Prot variant web pages*
  Missense mutation leading to single amino acid polymorphism (SAP) is the type of mutation most frequently related to human diseases. We have created Swiss-Prot Variant web pages to provide a summary of available sequence information as well as additional structural information on each variant. In particular, wherever possible, SAPs are modeled onto 3D protein structures and the users can visualize the models. The 3D models are updated with each weekly Swiss-Prot release. The Swiss-Prot variant pages are accessible from the NiceProt view of a Swiss-Prot entry (e.g. P06737) on the ExPASy server, via a hyperlink created for the stable and unique identifier FTId of each human SAP (e.g. VAR_007908).

- *Recent and forthcoming changes in Swiss-Prot*
  With Swiss-Prot release 41, we have introduced two documents to announce recent and forthcoming format changes in Swiss-Prot and TrEMBL.These documents replace the corresponding sections of the release notes, and contain detailed information about new keywords, new feature keys and comment topics, new cross-references and other format changes. Explicit links to new databases will no longer be announced here (i.e. under "What's new on ExPASy"), but in the document "Recent changes".

- *Implicit links*
  Implicit links (i.e. added on the fly to Swiss-Prot/TrEMBL entries when viewed with NiceProt) to the following databases have been added recently:
  - GenAtlas - A human gene database, e.g. P04406
  - HOBACGEN - Homologous bacterial genes database, e.g. P02937
  - HOVERGEN - Homologous vertebrate genes database, e.g. P02304
  - TAIR - The Arabidopsis Information Resource, e.g. Q38828
  - WorfDB - The C.elegans ORFeome cloning project, e.g. Q17330

- WormBase - Database on genetics, genomics and biology of C. elegans, e.g. Q17330

Information about the criteria for the creation of links to each of these databases can be found in the Swiss-Prot document List of databases cross-referenced in Swiss-Prot.

Whenever reference is made to the Transport Commission (TC) System in Swiss-Prot comments lines (CC), a link is added from the NiceProt view to the Transport Protein Database (example: P37905).

- *Visualization tool for peptide mass fingerprinting identification results*
  We have installed Biograph Applet v2.0, intended for the visualization of results of the PeptIdent, FindPept and FindMod tools. Links to Biograph are available as part of PeptIdent, FindPept and FindMod result pages.

March 21, 2003

New cross-references have been introduced in Swiss-Prot:

- Explicit links to GeneDb_SPombe, the Schizosaccharomyces pombe GeneDB, example: O94534.
- Implicit links to CleanEx, a gateway to public gene expression data via officially approved gene names, example: P02751.

There is a new Swiss-Prot document, arath.txt - Index of Arabidopsis thaliana entries and their corresponding gene designations.

An interface to PRATT has been implemented on ExPASy. PRATT is a tool to discover patterns that are conserved in a set of protein sequences. The patterns are reported using the PROSITE format. The ExPASy BLAST result representation has been modified to allow direct submission of a number of sequences to PRATT.

The ExPASy BLAST interface now allows to perform tblastn searches against individual microbial genomes (EMBL genome records, including plasmids).

Throughout the ExPASy server, the navigation bar at the top of every page now includes a search bar, for quick access to Swiss-Prot, TrEMBL, PROSITE, SWISS-2DPAGE, ENZYME, Taxonomy, HAMAP and ExPASy site search.

November 19, 2002

We are happy to announce a new **ExPASy mirror site in Bolivia**, http://bo.expasy.org, hosted by the Universidad Católica Boliviana in Cochabamba.

October 25, 2002

ExPASyBar, a very useful navigation bar to the most important databases and tools on the ExPASy server, has been developed by Martin Hassman, with input from the ExPASy team. ExPASyBar is an add-on to the Mozilla web browser (i.e. it does not work with Netscape, MS Internet Explorer and other browsers). Installation is very simple. ExPASyBar can be configured to use the ExPASy mirror of the user's choice (in the Edit/Preferences/Advanced/ExPASyBar menu of Mozilla).

August 27, 2002

The last "What's new on ExPASy" is more than a year old, which means that some of the changes and "new" features and services are not all that new anymore.... We are trying to list here the most important changes, and we will try to report new tools and documents again more frequently in the future!

=== ExPASy ===

- We are happy to announce that since the beginning of the year 2002, there is an **ExPASy mirror site in the USA**, http://us.expasy.org, hosted by the North Carolina Supercomputing Center (NCSC). Some users may have noticed upon their connection to www.expasy.org, that they are redirected to a mirror site that is closer to their geographic location, or that is less heavily loaded. If you feel that you are redirected to a mirror site for which the network connection is slow, please let us know .
- News on the FTP server:
  1. **PROSITE updated data** and documentation files are now made available via FTP even between releases, in the directory /databases/prosite/release_with_updates/. This data always corresponds to the version of the database available for web access via the PROSITE page.
  2. Up-to-date plain-text versions of all **SWISS-PROT documents** can be downloaded by ftp, in the directory /databases/swiss-prot/updated_doc/.
  3. Three "special selections" have been added:
     - merops.seq - all SWISS-PROT entries cross-referenced to the MEROPS database
     - mitoch.seq - Mitochondrion encoded proteins (entries with "Mitochondrion" on OG lines)
     - plastid.seq - Chloroplast and cyanelle encoded proteins (entries with "Chloroplast" or "Cyanelle" on OG lines)

=== TOOLS ===

Two tools have been added to our collection of sequence analysis and proteomics tools:

- The Sulfinator predicts **tyrosine sulfation sites** in protein sequences, based on Hidden Markov Models.
- PeptideCutter predicts potential **cleavage sites** cleaved by proteases or chemicals in a given protein sequence. It displays the query sequence with the possible cleavage sites mapped on it, as well as a table of cleavage site positions.

Major updates have been performed on two tools:

- The **ScanProsite** interface has been remodeled, with more options and databases, and a graphical view of the results. A standalone program, ps_scan is now available.
- The **BLAST** interface now allows searches in PDB. The output page displays hits found with Pfam HMMs and PROSITE profiles on the query sequence.

=== SWISS-PROT ===

- New **cross-references** have been *introduced* to various databases: AraC-XylS, Genew, Gramene, several 2D-PAGE databases, Ensembl, GeneLynx, ListiList, ModBase, PhosSite, ProtoNet, Source and TIGRFAMs.
  The List of databases cross-referenced in SWISS-PROT contains, for each database, a short description, the link type (explicit or implicit), and the server URL. In the case of explicit links, you can click on the word "Explicit" (example: Genew) to retrieve a list of all SWISS-PROT entries linked to the corresponding database.
  The cross-references to the following databases have been *deleted*, because the databases are either no longer available on the WWW, or because they have become commercial even for academic users: CarbBank, DOMO, GCRDb, Mendel, YEPD and YPD.

- The NiceProt view of SWISS-PROT has been further improved: access to documentation has been facilitated by adding "mouse-over" hypertext links from various sections in NiceProt to

the corresponding information in the user manual. Those hypertext links, which give access to documentation rather than the data related to the protein entry, are visually different from the ordinary hyperlinks. While they are not immediately recognizable as such, the user can see that they are clickable by moving the mouse pointer over the section headings such as "References" or "Keywords". A short description of the linked information appears at the bottom of the web browser, and when clicked, a small additional window is opened with related information extracted from the user manual.

Similarly, in the "Cross-references" section, the names of the databases to which an entry is cross-referenced are linked to the corresponding sections in the document dbxref.txt (List of databases cross-referenced in SWISS-PROT).

- Three SWISS-PROT documents have been released since the last announcement:
  - bucai.txt - Index of Buchnera aphidicola (subsp. Acyrthosiphon pisum) entries
  - mycpn.txt - Index of Mycoplasma pneumoniae strain M129 entries
  - plasmid.txt - List of plasmids

- The Human proteomics initiative (HPI) status report page has been remodeled and now contains more detailed information about the status of annotation of human SWISS-PROT entries.
- The HAMAP project aims to annotate semi-automatically **complete bacterial and archaeal proteomes** up to the quality standards of SWISS-PROT. Several proteomes have already been completed and are continuously updated. Up-to-date statistics are available on the HAMAP status page
- Note that upcoming format changes in the next SWISS-PROT release are always described in the release notes for the current release.
- Although not hosted physically on the ExPASy server, the NEWT Taxonomy browser is provided and maintained by members of the SWISS-PROT group, and serves as an entry point into SWISS-PROT and TrEMBL using taxonomic search criteria.

=== SWISS-2DPAGE ===

New **cross-references**, **reference maps**, and a **document** have been added:

- Cross-references to recent fully federated 2-DE databases, built with the Make2ddb package, are provided. These are now COMPLUYEAST-2DPAGE, PHCI-2DPAGE, PMMA-2DPAGE, and Siena-2DPAGE. The list of links is updated each time a SWISS-2DPAGE release is completed.
- SDS-PAGE and 2-D PAGE of nucleolar proteins from Human HeLa cells have been added to the list of reference maps . These masters are named respectively NUCLEOLI_HELA_1D_HUMAN and NUCLEOLI_HELA_2D_HUMAN.
- A FAQ (Frequently Asked Questions) has been provided. We hope you will find answers to most of your questions in this new document.

June 30, 2001

New cross-references have been added from relevant SWISS-PROT entries to three databases:

- SMART - Protein domain database (example: O43707).
- Leproma - Database dedicated to the analysis of the genome of the leprosy bacillus Mycobacterium leprae (example: Q9CBW4).
- MypuList - Mycoplasma pulmonis genome database (example: P58174).

The keyword "Complete proteome" has been introduced to all SWISS-PROT/TrEMBL entries

describing a protein which is thought to be expressed by an organism whose genome has been completely sequenced. This keyword is so far only used for microbial (bacterial and archaeal) proteins. A complete set of proteins from a microbial genome can therefore be obtained using this keyword across SWISS-PROT and TrEMBL.

A new and improved version of the NiceProt view of SWISS-PROT is available ( example). Some of its new features are:

- It provides a link to a printer-friendly view of a SWISS-PROT entry.
- It displays the length of certain features in the FT lines.
- It provides access to a new tool, the 'Feature aligner' which allows to select features for submission to the ClustalW multiple alignment program.

SWISS-PROT release statistics are now available for every update of the database. Among other parameters, statistics about database growth, average sequence lengths and amino acid composition, taxonomic origin, journal citations and database cross-references are presented, including some graphics.

A new view is available within the SRS Sequence Retrieval System. It displays, for each protein corresponding to a user query, gene name(s) and organism (in addition to the parameters ID, AC, description and sequence length which are displayed by the default view "Short description"). This new view is entitled "**Long description**" and is available from the menu "Use view ..." in the SRS query form.

The SIB Blast interface (accessible also via "Quick BLAST" or from the bottom of every SWISS-PROT/ TrEMBL entry) now offers the possibility to restrict the similarity search by using taxonomic criteria. A "Taxonomic View" of the results can also be obtained via the BLAST result page.

L'équipe Swiss-Prot a le plaisir de vous présenter le premier article de "Protéines à la Une", sa nouvelle rubrique de vulgarisation scientifique dédiée aux protéines qui font parler d'elles dans l'actualité.

January 18, 2001

- **SWISS-PROT**
  New cross-references have been added to three additional databases:
  - GlycoSuiteDB - a database of glycan structures; explicit links
    example: P00750
  - GeneCensus - a compilation of ORF data for the Saccharomyces genome; implicit links
    example: Q01802
  - HUGE - a database of human unidentified gene-encoded large proteins; implicit links
    example: P42330

- **NiceProt & SIB BLAST** The NiceProt view of SWISS-PROT/TrEMBL entries now contains a direct submission button requesting a blastp homology search of the protein against SWISS-PROT/TrEMBL/TrEMBLnew, on the SIB BLAST server ( *"Quick BlastP search"*). In the results of SIB BLAST searches on ExPASy (normal or "NiceBlast" output formats), the user can select a number of matching sequences and directly submit them to a ClustalW search, or retrieve and download the corresponding SWISS-PROT/TrEMBL entries.

- **Proteomics tools**
  - FindPept: This new tool can identify peptides that result from unspecific cleavage of proteins from their experimental masses, taking into account artefactual chemical modifications,

post-translational modifications (PTM) and protease autolytic cleavage.

- PeptIdent: Several new features have been added.
  - When searching SWISS-PROT, all alternative splice isoforms described in SWISS-PROT feature tables are included in the search (e.g. Isoform 12S of O43184).
  - New organism classes can be searched. For each of the available taxonomic available (e.g. Mammalia), a new section (e.g. other Mammalia) has been added, which comprises all entries not corresponding to any of the searchable subclasses (e.g. all Mammalia except human, bovine, rabbit, and other rodents).
  - For each matching protein in a PeptIdent result, buttons are available which allow further analysis of the protein by direct submission of the data to FindMod, FindPept, GlycoMod, PeptideMass and BioGraph.
- GlycoMod: Possible oligosaccharide structures suggested by GlycoMod are linked to the GlycoSuiteDB database of glycan structures, if they are reported in this database. The user can also select to display compositions reported in GlycoSuiteDB separately from the compositions not known in the database.

October 28, 2000

Several new features have been implemented on ExPASy during the last few months:

- The Swiss Center for Scientific Computing (CSCS) and the Swiss Institute of Bioinformatics provide a powerful and rapid new BLAST server. A submission form to this server is available from the bottom of each SWISS- PROT/TrEMBL entry on ExPASy. Results of blastp similarity searches submitted from this form are now parsed and displayed in a more user-friendly way, including a graphical representation and a link to NiceBlast. NiceBlast is a html table detailing complete descriptions of all matching proteins, including the full protein name, gene name, sequence length and organism.

- Sequences of alternatively spliced isoforms of the same protein are documented in the feature table of that protein sequence record. In collaboration with the SWISS-PROT group at EBI, a program varsplic.pl has been written to generate additional records from SWISS-PROT and TrEMBL, one for each splice isoform of each protein. The resulting data sets for SWISS-PROT and TrEMBL are available on our ftp server, along with a more detailed description of the project and information on how to obtain a local copy of the varsplic.pl program.

  The additional isoform entries have been added to the SWISS- PROT/TrEMBL databases underlying the BLAST server at SIB/CSCS Switzerland, and ScanProsite. Gradually, all other tools on ExPASy will be modified to handle splice isoforms. The NiceProt view of SWISS-PROT/TrEMBL provides links from the isoform name in the feature table (example: Q01432) to a page displaying the sequence of the corresponding isoform.

- In the framework of the HAMAP project, we provide clean non-redundant SWISS-PROT/TrEMBL data sets for all completely sequenced microbial genomes. These files are available on the ExPASy ftp server in SWISS-PROT and Fasta format, and can also be used for similarity searches on the SIB Blast server ("microbial proteomes").

  A **Genomic Proximity Viewer** is available for those microbial genomes where an ORF numbering system exists. For those organisms, it is possible to click on the ORF name in the SWISS-PROT/TrEMBL GN (gene) lines to obtain a list of proteins encoded by genes in proximity (example: P46448). The tool is also accessible from the HAMAP

complete proteome pages of those organisms. Example: Borrelia burgdorferi.

- The following cross-references have been added to relevant SWISS-PROT/TrEMBL entries:
  - InterPro - the Integrated Resource of Protein Families, Domains and Sites, integrating PROSITE, Pfam, PRINTS and ProDom. A link to a graphical view of the domain structure is also available; example: O15197.
  - MEROPS - a peptidase database; example: O96009.
  - NucleaRDB - a database of nuclear receptors (implicit links); example: O09018.
  - DIP - Database of Interacting Proteins (implicit links); example: P10275
- The Compute pI/Mw tool, if called for a list of proteins, can now produce, in addition to the usual verbose format, a table in text format that can be exported to an external application.
- Protein Spotlight is a periodical electronic review from the SWISS-PROT group. It is published on a monthly basis and consists of articles focused on particular proteins of interest. You can subscribe to receive each issue, free of charge, in HTML or PDF format.

April 26, 2000

**Proteomics Tools:**

- We are happy to announce a new tool in our suite of ExPASy protein identification and characterization tools:
  GlycoMod is a tool that can predict the possible oligosaccharide structures occurring on proteins from their experimentally determined masses. The program can be used for free or derivatized oligosaccharides and for glycopeptides. GlycoMod has been developed in collaboration with Nicolle Packer, initially at Macquarie University, Sydney, and later at Proteome Systems Ltd. GlycanMass is an associated tool which allows to calculate the mass of an oligosaccharide structure from its oligosaccharide composition.
- Detailed documentation is now available for the PeptIdent peptide mass fingerprinting identification tool.
- A number of new functionalities have been added to FindMod:
  - Results can now be obtained by email (as an alternative to receiving them on-line in the browser window), in form of an html file, with exactly the same functionality as for on-line display.
  - Several new enzymes have been added, mainly different versions of Chymotrypsin.
  - Results given in the "potential amino acid substitutions" table have been refined:
    - We no longer suggest amino acid (aa) substitutions occurring on the enzyme cleavage site and substituting the aa for an aa at which the enzyme does not cleave.
    - If the suggested aa substitution corresponds to a sequence variant or conflict as annotated in the SWISS-PROT feature table, this substitution is highlighted in color (green background for that table line), and a hypertext link is provided to the corresponding annotated variant or conflict.
- Compute pI/Mw can now be used with a file uploaded from the user's computer, if this file contains a list of SWISS-PROT/TrEMBL IDs/ACs.

**SWISS-PROT:**

- Dotlet, a diagonal dot-matrix program drawing a dotplot of two sequences, has been included in the set of tools that can be called directly from the bottom of each SWISS-PROT/TrEMBL entry on ExPASy. This allows to find repeats within the sequence.

- In the last few months, cross-references to the following databases have been added to relevant SWISS-PROT entries:
  - TubercuList - for entries from Mycobacterium tuberculosis
  - PRINTS - Protein fingerprint database
  - implicit links to BLOCKS - a database of multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins.

  Example entry: Q50705.

- There are 6 new SWISS-PROT documents:
  - humchr08.txt: Index of protein sequence entries encoded on human chromosome 8.
  - humchr09.txt:Index of protein sequence entries encoded on human chromosome 9.
  - humchr10.txt:Index of protein sequence entries encoded on human chromosome 10.
  - humchr11.txt:Index of protein sequence entries encoded on human chromosome 11.
  - dbxref.txt: List of databases cross-referenced in SWISS-PROT.
  - rprowaze.txt: Index of Rickettsia prowazekii strain Madrid E entries.

**ExPASy:**

- We are happy to announce a **new ExPASy mirror site**, at Peking University, China: http://expasy.pku.edu.cn/.
- We have completely revised the ExPASy server access statistics, which were previously frequently incomplete and erroneous. Every month, a table is updated which lists monthly access statistics for the main Swiss ExPASy server and for all our mirror sites.

October 4, 1999

- The ExPASy server has a **new mirror site** for North America, at the Canadian Bioinformatics Resource in Halifax, Canada. It can be reached at the URL http://expasy.cbr.nrc.ca/.

- The SWISS-PROT search by description tool has been extended to TrEMBL.
- There are five new SWISS-PROT documents:
  - humchr12.txt: an index of protein sequence entries encoded on human chromosome 12.
  - humchr14.txt: an index of protein sequence entries encoded on human chromosome 14.
  - humchr15.txt: an index of protein sequence entries encoded on human chromosome 15.
  - humchr16.txt: an index of protein sequence entries encoded on human chromosome 16.
  - annbioch.txt: SWISS-PROT annotation: how is biochemical information assigned to sequence entries

- When scanning a pattern against the SWISS-PROT/TrEMBL databases using the ScanProsite tool, users can now **restrict their searches** to an organism or a taxonomic range.
- The **NiceSite** view of PROSITE (example: PS00101) has been modified to include two new statistical values in its section of numerical results, namely
  **Precision** (true hits / (true hits + false positives)) and
  **Recall** (true hits / (true hits + false negatives)).

- A new parameter has been added to the list of parameters computed by the ProtParam tool: The program now calculates the **atomic composition** of a protein, in addition to molecular weight, theoretical pI, amino acid composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY).

June 16, 1999

The 'Nice' view tools for the databases provided on ExPASy (SWISS-PROT, SWISS-2DPAGE, PROSITE, ENZYME) have been developed in order to provide users with an easily readable

alternative to the original text file representation.

The following tools are available:

| Database | Tool | Example |
|---|---|---|
| SWISS-PROT | NiceProt | http://www.expasy.ch/cgi-bin/niceprot.pl?P14060 |
| SWISS-2DPAGE | Nice2DPage | http://www.expasy.ch/cgi-bin/nice2dpage.pl?P00938 |
| PROSITE | NiceSite | http://www.expasy.ch/cgi-bin/nicesite.pl?PS00661<br>http://www.expasy.ch/cgi-bin/nicedoc.pl?PDOC00566 |
| ENZYME | NiceZyme | http://www.expasy.ch/cgi-bin/nicezyme.pl?2.4.1.1 |

We have now changed all our tools and database search programs on ExPASy to display the 'Nice' version of a database entry by default. The programs displaying database entries in their original text formats continue to be maintained, and links are available from the 'Nice' views to the corresponding get-xxx-entry programs (e.g. get-sprot-entry).

If you maintain pages with links to entries from the above-mentioned databases, you might be interested to update these links to use the 'Nice' View if you prefer this representation to the original format. Otherwise you are, of course, completely free to keep the get-xxx-entry links.

May 24, 1999

- *Linking to ExPASy*
  We have revised the ExPASy file and directory structure, in order to have the vast amount of data that has accumulated on ExPASy since September 1993 available in a more structured manner, and to facilitate replication on our mirror sites. This has caused certain changes in html links, and we would like to ask our users to update their bookmarks and links accordingly. If in doubt, please refer to the document 'How to create html links to ExPASy'. At the same time we wish to reiterate our announcement of the ExPASy mirror sites in Taiwan and Australia. For your own convenience, please use the mirror site closest to you. Regular users might also bookmark the addresses of all ExPASy mirror sites to use as backup for the rare cases that their favourite ExPASy site is down or unreachable due to network problems.

  Please make sure to update all pointers using the old domain expasy.hcuge.ch, which was replaced by
  http://www.expasy.ch/ in March 1997 (!). The 'expasy.hcuge.ch' address might be disabled in the near future.

- *Protein identification tools*
  AACompIdent and MultiIdent have been revised, and the database choice has been extended to include TrEMBL. Results are now sent to the user in html format (rather than text only), and html links allow direct access to the matching SWISS-PROT/ TrEMBL entries.

- *SWISS-PROT cross-references*
  SWISS-PROT entries from Escherichia Coli entries with 'DR ECOGENE' lines are now directly linked to EcoGene at the University of Miami.
  There is a new type of cross-reference lines for sequence entries from Brachydanio rerio (Zebrafish): these entries are now linked to the Zebrafish Information Network (ZFIN) at the University of Oregon.

- *New features have been added to improve interactivity in accessing SWISS-2DPAGE*:

- All searching functions in the database can be accessed from the top page and results page of each keyword search function (example: search by description). This feature has been designed to facilitate the navigation between the different ways to query the database (by description, by access number, by authors, by full text search).
- A new tool is provided to retrieve in a table all the protein entries identified on a given reference map, with all 2-DE information: spot serial number, pI, Mw, mapping procedure, references (example).
- A new way to query the database is provided. From a user-entered amino acids sequence, one can display the estimated location on a choosen reference map ( example).

February 26, 1999

- *Several new features have been added to the PeptIdent peptide mass fingerprinting identification tool:*
  - It is now possible to search SWISS-PROT and/or **TrEMBL**.
  - In the page displaying the PeptIdent results, a button allows to perform a **new search** with slightly modified parameters by giving access to the PeptIdent form filled in with all previously used parameters.
  - For each matching protein, a direct link to BioGraph gives access to a **graphical representation** of the results of the PeptIdent query. BioGraph was developed by Daniel Doubrovkine and Anton Soudovtsev as a student project in the scope of the Bioinformatics course given at Geneva University.
  - The **sequence portion covered by the matching peptides** can optionally be displayed and highlighted in colour, as well as the difference between pI and Mw values of the matching proteins and the user-specified values.

- In the results of the SIM binary sequence alignment tool, a direct link has been addded to the PRSS program from EMBnet-CH which **evaluates the significance of a protein sequence similarity score**.

- Direct links have been added from the comments (CC) lines of relevant SWISS-PROT entries to the SWISS-PROT documents listing ribosomal protein families (e.g. RL2_ECOLI), aminoacyl-tRNA synthetases (e.g. SYC_HUMAN) and 7-transmembrane G-linked receptors (e.g. AA3R_MOUSE).

- Since the introduction of organism classification (OC) terms of the NCBI taxonomy with SWISS-PROT release 37, OS (organism species) lines have been linked to the corresponding pages of the NCBI taxonomy browser.

- The PROSITE full text search tool has been improved. Like in the SWISS-PROT/TrEMBL full text search program, wildcards can be used in query strings and search keywords can be combined with boolean operators.

- We have developed **Nice2DPage**, a tool that provides a user-friendly tabular view of SWISS-2DPAGE entries (example). The 'Nice2DPage View of SWISS-2DPAGE' is accessible from the top of each SWISS-2DPAGE entry on ExPASy.
- New hypertext cross-references have been added to SWISS-2DPAGE entries (e.g. P02997):
  - from the 2D comments lines (MAPPING, EXPRESSION LEVEL...), direct links have been added to the concerned citation in the SWISS-2DPAGE entry
  - from the 2D lines concerning AMINO ACID COMPOSITION and PEPTIDE MASSES data, direct links have been added to the concerned section in the user manual describing data format and protocols.

- Links to the Brenda enzyme database have been added to ENZYME entries.

October 27, 1998

- The SWISS-PROT/TrEMBL and SWISS-2DPAGE full text search tools have been improved. The databases are now indexed using the Glimpse search engine, wildcards can be used in query strings, more fields (line types) are indexed and response times are much shorter than before.

- We have developed **NiceProt**, a tool that provides a user-friendly tabular view of SWISS-PROT entries (example). The 'NiceProt View of SWISS-PROT' is accessible from the bottom of each SWISS-PROT entry on ExPASy.

- The following database cross-references and literature references have been added to SWISS-PROT entries on ExPASy:
  - DR links to the PRESAGE resource for structural genomics from Stanford University (e.g. P53878);
  - DR links from relevant immunoglobin entries to IMGT, the international ImMunoGeneTics database from the University of Montpellier (e.g. P01876);
  - References to the Worm Breeder's Gazette in the RL lines of relevant entries from *Caenorhabditis elegans* (e.g. Q09517).

- Users who wish to save and retrieve all SWISS-PROT entries originating from a species can do this via the SWISS-PROT document 'List of organism identification codes': By clicking on any of the species codes (e.g. DROME) and specifying a filename, one can save all corresponding entries to a file which can be retrieved from the anonymous ExPASy FTP server.

- The output format of the PeptIdent peptide mass fingerprinting identification tool has been improved. PeptIdent results now contain a table summarizing information about the matching proteins, from where the user can jump to the detailed listing for the corresponding peptides.

- The new experimental tool CombSearch provides a unified interface for simultaneous queries to several protein identification programs accessible on the web. CombSearch was written by Rémi Hammerli and Pavel Dobrokhotov as a student project in the scope of the Bioinformatics course given at Geneva University.

- A new page providing links to conferences and events is available and accessible from the ExPASy home page. If you know about any conferences on molecular biology or bioinformatics we encourage you to register .

- The ExPASy interfaces which allow the direct submission of a SWISS-PROT/TrEMBL sequence to BLAST servers at EMBnet-CH and NCBI have been modified to provide a more transparent selection menu of BLAST programs and databases. These programs are designed for similarity searches easily accessible from a SWISS-PROT/TrEMBL entry; for advanced searches with more options we recommend to use the original BLAST submission forms at EMBnet-CH or NCBI.

August 24, 1998

- *There is a new tool in our section 'Protein identification and characterization tools':* PeptIdent allows the identification of proteins using pI, Mw and peptide mass fingerprinting data. Experimentally measured, user-specified peptide masses are compared with the theoretical peptides calculated for all proteins in SWISS-PROT. A species (or group of species) can also be specified for the search. PeptIdent makes extensive use of the annotations

in SWISS-PROT and takes into account post-translational modifications as documented in SWISS-PROT.
Results are displayed on-line or can be sent by email, in form of a html table. The result file contains direct links to FindMod to further characterize matching proteins by predicting potential protein post-translational modifications and finding potential single amino acid substitutions, and to PeptideMass.

- There is a new document describing how to create HTML links to services on ExPASy.
- In July 1998, SWISS-PROT, PROSITE and ENZYME have undergone major releases.
- New hypertext cross-references have been added to SWISS-PROT entries (example: P98073):
  - in RX lines: Medline abstracts corresponding to SWISS-PROT references can now also be consulted at the Weizmann Institute of Science in Israel, in addition to the archives at NCBI, ExPASy and GenomeNet Japan. These links have also been added to SWISS-2DPAGE entries.
  - DR DOMO lines have been added: These links provide direct access to relevant information in the DOMO database of homologous protein domains maintained by Jérôme Gracy at Infobiogen.
  - At the bottom of the page displaying a SWISS-PROT/TrEMBL entry, there are now direct links for submission of the sequence to ScanProsite and ProfileScan.
  - RL lines: Relevant SWISS-PROT entries are now directly linked to the Plant Gene Register, an electronic publication for articles describing the isolation and DNA sequence determination of plant genes (example: P48422).
  - The ExPASy interface to the BLAST server at EMBnet-CH now uses their new BLAST2 client, replacing WU-BLAST.

June 13, 1998

- The ExPASy server presents itself in a new layout: the home page, database entry pages, the tools page and many other pages have been redesigned for easier navigation and better readability.

  Users can now also use (in addition to the home page and ExPASy Index) the newly created clickable ExPASy site map to find useful tools, documents and services available on our server, and to find out about functional links between them.

  A new documentation page has been created which presents a complete table of documents available on ExPASy.

- There are two new SWISS-PROT documents:
  - humpvar.txt: an index of human proteins with sequence variants
  - humchr17.txt: an index of protein sequence entries encoded on human chromosome 17.
- Protein domains, chains etc. documented in the SWISS-PROT feature tables, if corresponding to subsequences of at least 10 amino acids, can now be directly submitted to a BLAST similarity search from the pages highlighting these subsequences. Example: DOMAIN EXTRACELLULAR ALPHA-1 (1A24_HUMAN).
- Two bugs have been corrected in ExPASy tools:
  - There was a small error in the computation of extinction coefficients by ProtParam: The contribution of Cysteines to the extinction coefficient (Gill S.C., von Hippel P.H. Anal: Biochem. 182:319-326(1989)) of a protein is only half of the values used previously in ProtParam, which results in slightly different values for the extinction coefficient.

- Our Translate tool no longer ignores base-ambiguity characters such as M, W, Y, etc. Previously performed translations for DNA sequences containing characters other than A,C,T,U,G, and N are likely to have been incorrect.

We apologize for any inconvenience caused by these errors and encourage our users to continue to send us their comments and bug reports.

March 27, 1998

- *There is a new tool in our section 'Protein identification and characterization tools':* FindMod is a program for the de novo discovery of protein post-translational modifications. It examines peptide mass fingerprinting results of known proteins for the presence of currently 18 types of PTMs of discrete mass. This is done by looking at mass differences between experimentally determined peptide masses and theoretical peptide masses calculated from a specified protein sequence. If a mass difference corresponds to a known PTM not already annotated in SWISS-PROT, "intelligent" rules are applied that examine the sequence of the peptide of interest and make predictions as to what amino acid in the peptide is likely to carry the modification.

- *Improved tools:*
  PeptideMass, which calculates masses of peptides and their posttranslational modifications for a given protein sequence, can now consider up to 3 missed cleavages. Post-translational modifications may be specified for a sequence in raw sequence format, and substitution tables are available to simplify the interpretation of the results for peptides concerned by database conflicts, variants or splicing variants.

  TagIdent can now search in SWISS-PROT, TrEMBL or both databases. It is also possible to perform an additional scan of a short sequence tag against all fragments contained in the database(s), even if pI and Mw cannot be computed for these proteins.

  MultiIdent (identification using pI, MW, amino acid composition, sequence tag and peptide mass fingerprinting data) is available for constellation 2 (Ala, Ile, Pro, Val, Arg, Leu, Ser, Asx, Lys, Thr, Glx, Gly, Met, His, Phe and Tyr. (Asp+Asn=Asx; Gln+Glu=Glx; Cys and Trp are not considered)) and constellation 4 (like constellation 2, but Gly is not considered).

- Several months ago, we started to distribute and update weekly, a set of data files that can be used to build a non-redundant protein sequence database consisting of SWISS-PROT, TrEMBL and TrEMBL updates. There is now a document explaining the contents and principles of this database.

- Information about the current release and update status of SWISS-PROT has been added to the SWISS-PROT page (currently 'Release 35 and updates up to 20-Mar-1998: 71198 entries').

- New hypertext cross-references have been added to SWISS-PROT entries:
  - in RX lines: Medline abstracts corresponding to SWISS-PROT references can now also be consulted on the Japanese GenomeNet server in addition to the archives at NCBI and ExPASy.
  - in DR PDB lines: Local copies of PDB entries are available. The user is now given the choice between accessing 3D structure information (e.g. 2hhe) in Geneva or Brookhaven (US). Both links provide direct access to 3D structure information in various formats, as well as hypertext links to servers offering related information.
  - DR PROTOMAP lines have been added: These links provide, for a SWISS-PROT entry, a cluster (group) of related proteins as classified by the ProtoMap server at Hebrew University, Jerusalem.

Example: DEFN_HUMAN.

- SWISS-2DPAGE is now available to be searched by the SRS Sequence Retrieval System.

February 13, 1998

- The SWISS-PROT full text search tool has been redesigned and improved. Boolean operators (AND, OR, NOT) can be used to combine and restrict queries, and special characters such as _ - # ' ( ) , . / are allowed as part of words (as used in SWISS-PROT).

- SWISS-PROT author names (RA lines) have been linked to a page listing all SWISS-PROT entries which contain references to articles (co-) authored by this author.

- The ExPASy interface to the EMBNet-CH BLAST server now contains a new option: This BLAST process manages two job queues: a (presumably) fast one and a slow one. Based on the sequence provided and the database requested, the process makes an (educated ???) guess to decide if the query will require more than 5 minutes of CPU time. Small jobs are allowed to proceed in the fast queue, while the others are forced to the slower one. If an e-mail address is provided, results of slow jobs will be automatically mailed back, while fast jobs will proceed as before.

- Two features have been added in SWISS-2DPAGE to facilitate visualisation and differentiation of spots:
  - If you click on a spot in one of the SWISS-2DPAGE maps (e.g. Plasma), the '2D' line describing this spot in the corresponding SWISS-2DPAGE entry is highlighted in green.
  - Hypertext links have been added from spot serial numbers on SWISS-2DPAGE '2D' lines to the master image for the protein, in which the spot with this serial number is highlighted in green (in contrast to the other spots displayed in red). Example: P00450.

January 13, 1998

- Since November 1997, SWISS-PROT, PROSITE, ENZYME and SWISS-2DPAGE have all gone through major releases.

- There is a new program that allows you to randomly retrieve a SWISS-PROT or TrEMBL entry.

- A new output format option has been added to our Translate tool. When translating a nucleotide sequence into a protein sequence, you can now also select to include, for each of the six open reading frames, the nucleotide sequence in the output.

- Cross-references and direct links to the Mendel Plant Gene Nomenclature Database have been added in corresponding SWISS-PROT entries. Example: P12084. There also is a file containing all SWISS-PROT entries with cross-references to Mendel in our series of "special selections", which is updated weekly and can be downloaded from our anonymous FTP server.

- Proteins which are documented to belong to an uncharacterized protein family in the SWISS-PROT CC (comments) lines, have been linked to the SWISS-PROT document upflist.txt. Example: P55061.

November 27, 1997

- In SRS (Sequence Retrieval System), SWISS-PROT DR (Database crossReference) and RC (Reference Comment) lines have been indexed. You may search for e.g. all entries with cross references to PDB (enter 'PDB' in the DbName field), or all proteins that have been found in E.coli strain K12 (enter 'K12' in the 'RefComment' field).

- It is now possible to retrieve a number of SWISS-PROT/TrEMBL entries by specifying a list of accession numbers or entry names (ID).

- There are 4 new SWISS-PROT documents:
  - humchr18.txt: an index of protein sequence entries encoded on human chromosome 18;
  - pcc6803.txt: an index of Synechocystis strain PCC 6803 entries;
  - deleteac.txt: an index of deleted accession numbers.
  - upflist.txt: UPF (Uncharacterized Protein Families) list and index of members.

October 7, 1997

- We have implemented a search index, ExPASy Index, to help you find information within the ExPASy server. The index contains all the documents of ExPASy (currently about 800), except the database entries. It has been automatically indexed by the Marvin robot.
  Our new service BioHunt uses the same concept and allows you to search the internet for molecular biology information. In the current version, 17136 documents have been indexed.

- The ScanProsite tool has been modified to work with TrEMBL as well as SWISS-PROT. Furthermore, the part of the program which allows to scan a pattern against SWISS-PROT (and TrEMBL) has been improved and now avoids the previously frequent 'Document contains no data' error for large scan results.

- In PeptideMass, the set of post-translational modifications with discrete mass differences considered in peptide mass computation now also contains O-GlcNac (documented as `FT CARBOHYD GLCNAC` in SWISS-PROT) and C-Mannosylation of Tryptophan (`FT CARBOHYD C-MANNOSYL`). Thus, 17 post-translational modifications are now considered in PeptideMass. For examples, try CRAA_BOVIN or RNKD_HUMAN, don't forget to select "display all known post-translational modifications" and click on the "Perform" button.

- There is a new SWISS-PROT document:
  mgdtosp.txt - Index of MGD entries referenced in SWISS-PROT.

- Hyperlinks have been added from SWISS-PROT entries to the TIGR Microbial Database, which provides links to the information provided by TIGR on the genes encoded in the genomes they have sequenced (so far these are: Haemophilus influenzae, Helicobacter pylori, Methanococcus jannaschii, and Mycoplasma genitalium). (Example: FDHB_METJA)
  We have also created a specific file containing all SWISS-PROT entries containing cross-references to the TIGR database in our series of "special selections", which is updated weekly.

- SWISS-PROT reference (RL) lines and PROSITE references referring to one of the journals available at IDEAL, an online electronic library containing all 175 Academic Press journals, now contain direct links to the IDEAL server if the article was published in 1996 or later. From this, a 'Guest login' leads to the abstract of the article. (Example: RGSE_RAT)

September 5, 1997

Some new features of ExPASy:

- The PeptideMass program has been modified to take into account up to 2 missed cleavage sites. A new column 'MC' has been added to the output which indicates the number of missed cleavages, and peptides resulting from 0, 1 or 2 missed cleavages are displayed in different colours.
- A new parameter has been added in the ProtParam program: ProtParam results now include the grand average of hydropathicity (GRAVY) for a given protein.

- At the bottom of each SWISS-PROT and TrEMBL entry, there is now a link to a page displaying the entry in FASTA format (example: P11553).
- The local submission form to the WU-BLAST server at Lausanne has been changed to use as the default database the set of non-redundant protein databases SWISS-PROT, TrEMBL and TrEMBL_NEW.
- There are two new SWISS-PROT documents:
  - metallo.txt - Classification of metallothioneins and index of MT entries
  - hpylori.txt - Index of Helicobacter pylori strain 26695 chromosomal entries
- The display of current and previous Swiss-Flash bulletins has been redesigned: A table is available which lists all Swiss-Flash bulletins by category, including date, title and author of the bulletins.

July 24, 1997

We have now an SRS server (version 5) running on ExPASy. SRS (Sequence Retrieval System) allows you to retrieve entries across multiple databases with more sophisticated criteria than those allowed by the text-search interfaces available from the SWISS-PROT top page.

You can combine all the fields with logical operators and achieve queries like:

- Give me all vertebrate proteins having a PH domain and that are longer than 1000AA or
- Give me all calcium-binding proteins localized in the endoplasmic reticulum.

Five databases are indexed: SWISS-PROT, TrEMBL, TrEMBL_NEW, PROSITE, and ENZYME. SWISS-PROT and TrEMBL are updated on a weekly basis so that the set of these two databases stays non-redundant.

TrEMBL entries are now fully accessible on ExPASy via a cgi-script. The hypertext version of TrEMBL contains links to various databases and allows direct access to sequence analysis tools such as Swiss-Model, Blast, ProtParam, ProtScale, Compute pI/Mw and PeptideMass, as is the case for SWISS-PROT.

If you wish to link to a TrEMBL entry, you can use the following URL:

```
http://www.expasy.ch/cgi-bin/get-sprot-entry?<TrEMBL-AC>
```

e.g. to create a link to TrEMBL entry Q00061, use:

```
http://www.expasy.ch/cgi-bin/get-sprot-entry?Q00061
```

June 6, 1997

We are actively seeking any type of updates and/or corrections of SWISS-PROT entries, whether they have been published or not, and we encourage our users to submit us their suggested updates or corrections. This can be done using our new submission form, which can be accessed through an active link from the SWISS-PROT home page or from the bottom of each SWISS-PROT entry. Please read the tips and guidelines to find out what type of information we are seeking and how to proceed. We would already like to thank our users in advance for any contribution they can make in updating and correcting SWISS-PROT!

The tool which allows you to visualize and highlight the subsequence corresponding to a line in a SWISS-PROT feature table (FT) has been improved and is now using colour to highlight the subsequences in question. Example: in FA9_HUMAN:

```
FT   DOMAIN       93    129       EGF-LIKE 1, CALCIUM-BINDING (POTENTIAL).
```

May 21, 1997

At the bottom of each page displaying a SWISS-PROT entry, you will now find a link to a graphical Feature Table viewer (Java Applet) written by Thomas Junier at the Bioinformatics Group of ISREC Lausanne.

We have added several new hyperlinks in SWISS-PROT entries:

- The DR lines containing cross-references to EMBL/GenBank/DDBJ now include a link to a page displaying exclusively the corresponding CoDing Sequence (CDS).
- The RL lines referring to recent articles in certain journals whose WWW servers are maintained in collaboration with HighWire Press are now active hyperlinks to the abstracts of the corresponding articles. From the abstract page you can frequently access directly a full text on-line version of the article. The journals include J. Biol. Chem., Proc. Natl. Acad. Sci. USA, Science, Cell, etc.
- Entries with cross-references to MIM are now also linked (through a new virtual " `DR GeneCards`" line) to GeneCards, a database integrating information about the functions of human genes and their products, and of biomedical applications based on this knowledge. Example: BRC1_HUMAN.
- Entries belonging to family 1 of G-protein coupled receptors (as documented in feature tables) now contain active links to GPCRDB-Snakes diagrams (through the new virtual " `DR GPCRDB-Snakes`" line) prepared by the GPCRDB group at EMBL Heidelberg. Example: 5H1A_HUMAN.

There are 3 new SWISS-PROT documents:

- humchr19.txt: an index of protein sequence entries encoded on human chromosome 19
- ngr234.txt: a table of putative genes in Rhizobium plasmid pNGR234a
- initfact.txt: a list of translation initiation factors

On the ExPASy anonymous FTP server, the SWISS-PROT update files new_seq.dat, upd_ann.dat and upd_seq.dat are now also available in compressed form in the directory /ftp/databases/swiss-prot/updates_compressed/.

March 27, 1997

We have modified and improved access from ExPASy to various BLAST (Basic Local Alignment Search Tool) similarity search services:

In the tools page, you can now choose between 5 different interfaces to BLAST servers in Switzerland, the USA and Germany:

Switzerland:

Running on a 2-processor Pentium Pro machine, the new WU-BLAST server at EMBNet Switzerland in Lausanne has a faster response time than the EPFL server, and should be more stable. As opposed to the original NCBI BLAST algorithm, WU-BLAST generates gapped alignments. A full set of weekly updated databases is provided.

- Local interface to WU-BLAST at EMBNet-CH (Lausanne)
- Original interface to WU-BLAST at EMBNet-CH (Lausanne)

USA:

- Local interface to BLAST at NCBI
- Original interface of BLAST at NCBI

Germany:

- WU-BLAST at Bork's group in EMBL (Heidelberg)

For direct BLAST submission from a SWISS-PROT entry (icons at the bottom of the page displaying

an entry - example), you have the choice between the servers at NCBI and EMBNet-CH.

The following documents have been added to the list of SWISS-PROT documents:

- bloodgrp.txt - Blood group antigen proteins
- fly.txt - Index of Drosophila entries and their corresponding FlyBase cross-references
- mjannasc.txt - Index of Methanococcus jannaschii entries
- mgenital.txt - Index of Mycoplasma genitalium strain G-37 chromosomal entries

March 17, 1997

We have completely rewritten the Swiss-Shop sequence alerting system for SWISS-PROT that allows you to automatically obtain (by email) new sequence entries relevant to your field(s) of interest.

In the new version of Swiss-Shop, some new features have been added:

As before, you can either launch a sequence/pattern based search or a keyword based search.

- For a sequence based search, you need to specify a SWISS-PROT ID or AC or a raw protein sequence, and your sequence will be scanned, at each weekly update of SWISS-PROT, against the new sequences in the database using the alignment program BLAST. Sequences thus found to be similar to your protein will be sent to you by email. It is up to you to specify the BLAST probability threshold for P(N) (the probability that the alignment is real and not random), and you will receive a list of all sequences for which this probability is below the specified value.
- For a pattern based search, enter a PROSITE ID or AC or a pattern in PROSITE format, and Swiss-Shop will scan this pattern, at each weekly update of SWISS-PROT, against the sequences that have been added in SWISS-PROT since the last weekly update. You will receive the list of new entries matching your pattern.
- For a keyword based search, it was previously possible to specify keywords from SWISS-PROT OS, OC, OG (taxonomy), RA (authors), KW, DE, CC lines. In addition to these lines, you can now also search DR (Cross-references to other databases) and FT (feature) lines with one or more specified keywords. Swiss-Shop will look for these keywords on the corresponding lines of all SWISS-PROT entries added in the database since the last weekly release.

Furthermore, we now offer you 4 different output formats. You can choose to receive the sequences matching your query

- as a file in SWISS-PROT format or
- as a list of SWISS-PROT accession numbers or
- in form of a short report containing information from SWISS-PROT ID, AC, DE, OS lines or
- as a list of SWISS-PROT accession numbers with hypertext links to the corresponding entries on the ExPASy WWW server. This allows you to view your email message with your Web browser and to follow the hypertext links to the full entries on ExPASy.

You can further specify if you wish to be notified every time Swiss-Shop is run, even if there are no new sequences matching your query, or to receive an email report only when there are new SWISS-PROT entries matching your search terms.
You can specify the expiration date of your request, the default being one year after submission.
For editing previous requests (e.g. to update the expiration date or to modify search criteria) you can enter a password for each new request. This allows you to open the request later and edit it on-line rather than deleting it and submitting a new one.

March 6, 1997

New and improved protein identification tools:

*There is a new tool on ExPASy:*

- MultiIdent: This tool achieves protein identification using parameters such as protein species, estimated pI and MW, AA composition, sequence tag, and peptide mass fingerprinting data. It is particularly suited to the identification of proteins across species boundaries. Currently, the program works by first generating a set of proteins in the database with AA compositions close to the unknown protein, as for AACompIdent. Theoretical peptide masses from the proteins in this set are then matched with the peptide masses of the unknown protein to find the number of peptides in common (number of "hits"). Three types of lists are produced in the results. Firstly, a list where proteins from the database are ranked according to their AA composition score; secondly, a list where proteins are ranked according to the number of peptide hits they showed with the unknown protein; and thirdly, a list that shows only proteins that were present in both the above lists, where these proteins are ranked according to an integrated AA and peptide hit score. In all these lists, protein pI, MW, and species of origin (using a term from SWISS-PROT OS or OC lines) and keywords can be used, as in AACompIdent, to increase the specificity of searches.

*The following tools have been improved, offering numerous additional features:*

- AACompIdent (identification of a protein from its amino acid composition)
  You can restrict your search by specifying one or more term(s) from the OS or OC lines of SWISS-PROT (example: *HOMO SAPIENS* or *MAMMALIA*). You can also enter a keyword appearing on the KW lines of SWISS-PROT to further restrict your search. For example, a keyword of "CALCIUM-BINDING" could be used in conjunction with the OC term "MAMMALIA" to see if a user- entered protein matches well with any mammalian calcium-binding proteins in the database.
- TagIdent now allows, for one or more species (term from SWISS-PROT OS or OC lines) and with an optional keyword,
  1. the generation of a list of proteins close to a given pI and Mw,
  2. the identification of proteins by matching a short sequence tag of up to 6 amino acids against proteins in the SWISS-PROT database close to a given pI and Mw,
  3. the identification of proteins by their mass, if this mass has been determined by mass spectrometric techniques.

  For PeptideMass, Compute pI/Mw, AACompSim and all the above-mentioned tools, documentation and references have been added and the submission forms have been reformatted and improved.

March 4, 1997

Thanks to the generosity of the Geneva Government, we have been able to acquire a new computer for the ExPASy server (a Sun Microsystems Ultra Server Enterprise 2). The server is now accessible at URL:

http://www.expasy.ch

The old URL remains valid for some time.

January 9, 1997

Some new features of ExPASy:

- New active links have been established from SWISS-PROT entries
  - to the TRANSFAC database of transcription factors;
  - from Bacillus subtilis entries to Micado (MICrobial Advanced Database Organization) at

INRA, France;
  - to local copies of MEDLINE abstracts. We now give the user the choice of retrieving a MEDLINE abstract (example: 90368558) from either NCBI or Geneva;
  - to our Peptide Mass tool which cuts a protein sequence with a chosen enzyme and computes the masses of the received peptides.
- From Release 35 on, SWISS-PROT comments (CC) lines can contain a new 'topic' "DATABASE", which contains information about related databases catering for a specific protein or a for a very limited number of proteins. Most of these databases are mutation databases, reporting defects linked to a genetic disease. If such a database is available electronically, the CC DATABASE lines provide the relevant electronic coordinates, e.g. in P29965 (CD4L_HUMAN):

```
CC   -!- DATABASE: NAME=CD40Lbase; NOTE=European CD40L defect database;
CC       WWW="HTTP://www.expasy.ch/cd40lbase/";
CC       FTP="ftp.expasy.ch/databases/cd40lbase".
```

- There is a new SWISS-PROT document:
  yeast13.txt - a list of Yeast Chromosome XIII entries.
- Two new features have been added in ENZYME entries:
  - direct links from an enzyme to all relevant maps of Boehringer Mannheim's Biochemical Pathways and
  - links to the WIT (What Is There) database of metabolic pathways.

November 26, 1996

The Boehringer Mannheim Biochemical Pathways maps and index have been digitised and are now accesible on this server. Enter a keyword (such as, for example *Oxoacyl*) and surf on the biochemical pathways maps.

November 11, 1996

CD40Lbase, The European CD40L Defect Database prepared by Manuel Peitsch, has been made accessible through this server. The purpose of CD40Lbase is to collect clinical and molecular data on CD40 ligand defects leading to X-linked Hyper-IgM syndrome.

A new tool is available from the Tools page: The PeptideMass Peptide Characterisation Software. This program is designed to calculate the theoretical masses of peptides generated by the chemical or enzymatic cleavage of proteins, to assist in the interpretation of peptide mass fingerprinting and peptide mapping experiments. Protein sequences can be provided by the user or can be a code name for a protein in the SWISS-PROT protein database. When proteins of interest are specified from SWISS-PROT, the program considers all annotations for that protein in the database, and uses these in order to generate the correct peptide masses and warn users about peptides that are not likely to be found when undertaking peptide mass fingerprinting. Many protein post-translational modifications which affect the masses of peptides can thus be taken into consideration.

In PROSITE and Enzyme, we have added the possibility to save all referenced SWISS-PROT entries to a file on our anonymous FTP server (in the outgoing directory).

The Compute pI/Mw tool has been included in the list of sequence analysis tools that can be directly accessed from a SWISS-PROT entry.

Two new SWISS-PROT documents are available:
- humchr20.txt - an index of protein sequence entries encoded on human chromosome 20
- tisslist.txt - a list of the currently valid values for the "TISSUE" topic of the RC line type in SWISS-PROT.

September 30, 1996

A new SWISS-PROT document has been added: ribosomp.txt - an index of ribosomal proteins classified by families on the basis of sequence similarities.

In ec2dtosp.txt, an index of E. coli Gene-protein database (ECO2DBASE) entries referenced in SWISS-PROT, we have established direct links to ECO2DBASE, and SWISS-PROT entries now also contain links to ECO2DBASE.

At the end of each page displaying a SWISS-PROT entry we have added links to our sequence analysis tools ProtParam and ProtScale,which allows the user to directly submit the SWISS-PROT sequence to these tools.

September 19, 1996

Some new features of ExPASy:

- We have created a new protein identification tool called TagIdent. This is a modification of the old tool **GuessProt**. The user can now identify proteins from 2-D gels by giving protein pI and MW estimates, a species or organism classification of interest, and a short sequence tag of up to 6 amino acids. This tag can be derived from the N-terminus, the C-terminus or from internal peptides of a protein. The results are now sent to the user by e-mail, allowing many searches to be done at the same time. If you only want to generate a list of potential proteins in a specific pI or MW range (as was the function of the old tool GuessProt), do not select the TAG option in the form.
- An email option has been added to the tool ScanProsite: if you want to scan a pattern against SWISS-PROT, you have now the option of having sent the results of your query by email, which should avoid previously frequent timeout problems and is particularly useful for complex patterns.
  ScanProsite, which only scans SWISS-PROT with PROSITE *pattern* entries (as opposed to *rule* and *matrix* entries), can now also be used with the PROSITE rule entry PS00013, PROKAR_LIPOPROTEIN.
- SWISS-PROT entries have been linked to DDBJ, the DNA Data Bank of Japan. We have also added direct links to the Bacillus subtilis genomic data bank, SubtiList and to the Yeast Protein Database YPD to relevant SWISS-PROT entries.
- Links have been established from most feature (FT) lines of SWISS-PROT entries to pages that highlight the subsequence in question, both in 1- and in 3-letter amino acid codes. Example: in FA9_HUMAN:

```
FT   DOMAIN       93     129       EGF-LIKE 1, CALCIUM-BINDING (POTENTIAL).
```

- We have added three new SWISS-PROT documents:
  humchrx.txt - an index of protein sequence entries encoded on human chromosome X
  yeast7.txt - a list of Yeast Chromosome VII entries
  yeast14.txt - a list of Yeast Chromosome XIV entries.
- 2D Hunt, a database created and continuously updated by the Marvin robot contains sites related to electrophoresis and specifically to 2-D electrophoresis. It is now searchable from the SWISS-2DPAGE top page.

April 11, 1996

AACompIdent: New options - AACompIdent is a tool which allows the identification of a protein from its amino acid composition. It searches SWISS-PROT for proteins, whose amino acid compositions are closest to the amino acid composition given. Two new options and a new constellation have been added to this tool:

**A. C-Terminal display in tagging option**

The user may now choose between displaying the C or N terminal side of the proteins that score best.

**B. Permutation search in tagging option**

This option searches for all permutations of the given tag in the sequences.

**C. Constellation 4**

Constellation 4 has been added: Ala, Ile, Pro, Val, Arg, Leu, Ser, Asx, Lys, Thr, Glx, Met, His, Phe and Tyr. (Asp+Asn=Asx; Gln+Glu=Glx; Gly, Cys and Trp are not considered).

March 22, 1996

We have added a new tool, ProtScale which allows you to compute and represent the profile produced by an amino acid scale on a selected protein. 50 scales are provided, including 'classics' such as the Kyte and Doolittle hydrophobicity scale.

Links have been added between relevant SWISS-PROT entries and the 2D gel protein databases at Harefield.

A new SWISS-PROT document has been added which describes the nomenclature of glycosyl hydrolases (GH) and that includes an index of sequences that belong to the various GH families.

A PC (MS-Windows) version of LALNVIEW (graphical viewer for pairwise alignments) is now available.

Nicolas Guex has produced a new logo for PROSITE.

February 16, 1996

We have added a new tool, SIM which computes a user defined number of best non-intersecting alignments between two sequences. The results of the alignment can be viewed graphically using the LALNVIEW program developed by Laurent Duret and which is currently available for Macs and UNIX.

Additional links have been added in the tools page, notably to the Weizmann Institute ultra-fast rigorous (Smith/Waterman) similarity searches using the Bioccelerator and to the Garnier, Osgoodthorpe and Robson (GOR) secondary structure prediction method at SBDS.

The SeqAnalRef database now includes a section listing author's email and eventually also WWW home pages. It is also possible to access the links from a page displaying either a reference list or a single reference.

Amos has recently started to create a list of Biomolecular servers for his own usage, but as some people have asked to access this list (which is under construction), we are making it available from the ExPASy top page. Many other small changes were carried out in the last two months.

We thank you for using ExPASy (we have now reached a cumulative total of 4 million connections).

December 14, 1995

After 29 months of existence the ExPASy molecular biology server received a new logo, designed and produced by Nicolas Guex.

October 23, 1995

The Melanie page has been reorganised. With the announcement of release 2.1 of the **Melanie II 2-D PAGE analysis software package**, a complete up-to-date description of the software as well as a comprehensive tutorial are now available.

October 13, 1995

Links have been added between SWISS-PROT Escherichia coli K12 chromosomal entries and the EcoCyc database, the encyclopedia of E. coli Gene and Metabolism.

You can now seach in PROSITE by citation.

October 9, 1995

Some new features of ExPASy:

- Search in SWISS-PROT by citation - When you call this option, you are prompted to enter the name of a journal and optionally a volume number and/or a year. The program is written in such a way that you can enter either the full name of a journal or its official abbreviation.
- RandSeq - a new tool to generate random protein sequences.
- SWISS-PROT document haeinflu.txt - Index of Haemophilus influenzae RD chromosomal entries and gene names with links to the TIGR and EMBL servers.
- SWISS-PROT document submit.txt - Description of how to submit sequence data to the SWISS-PROT data bank.
- SWISS-PROT document aatrnasy.txt - List of aminoacyl tRNA synthetases.
- Swiss-Jokes - A new page to give access to our collection of jokes from the fields of molecular biology and of computing.

Many other changes have been done, such as the redesign of the Geneva local pages; the addition, in the tool page, of a link to ProfileScan.

It should also be noted that when you search in SWISS-PROT by either description or by full text and that your seach criteria returns more than two entries, you can save these entries to a file on our anonymous FTP server (in the outgoing directory).

September 19, 1995

AACompIdent: New options - AACompIdent is a tool which allows the identification of a protein from its amino acid composition. It searches SWISS-PROT for proteins, whose amino acid compositions are closest to the amino acid composition given. A new option and a new constellation have been added to this tool:

**A. Tagging option**

With this option, the first 40 amino acid of each protein are printed in the result, instead of the protein name. One may optionnally also enter a tag (a short seuqnece, typically 3 to 8 residues) which will be matched with the sequences of the resulting proteins. Proteins matching the tag will be marked.

**B. Free constellation**

This is a free constellation, that is one may select any amino acid constellation he/she likes. One just have to fill in the composition values for the selected amino acids. The values will then be normalised, so that the total make 100 (percent).

September 4, 1995

A new page has been created: WORLD-2DPAGE is an index to all known federated 2-D PAGE database servers, as well as to 2-D PAGE related servers and services.

July 22, 1995

A new tool has been implemented on ExPASy, ProtParam allows the computation of various physical and chemical parameters for a given protein stored in SWISS-PROT or for a user entered sequence. The computed parameters include the molecular weight, theoretical pI, amino acid composition, extinction coefficient, estimated half-life, instability index and aliphatic index

The Journal of Biological Chemistry (JBC) has a WWW server where abstracts and full text of articles are made available. We are happy to announce the implementation of what we believe to be the first direct link in a sequence database between a reference and the full text version of a cited article. Recent JBC references are directly linked to the corresponding entry point in the JBC server. If you want to see such a link, take a look at reference 3 in SWISS-PROT entry KDSA_ECOLI.

The SWISS-PROT document file jourlist.txt which provides information on all the journals cited in that database, now contains links to WWW or Gopher servers set up by a variety of publishers of academic journals.

Two new SWISS-PROT document have been added, one is a nomenclature and index of peptidase sequences, the other is the list of Yeast Chromosome VI entries in SWISS-PROT

June 19, 1995

A new tool has been implemented on ExPASy, ScanProsite allows to either scan a protein sequence the occurence of patterns stored in the PROSITE database or to scan the SWISS-PROT database - including weekly releases - for the occurence of a pattern.

We are happy to announce a new ""service"" Swiss-Quiz The principle of this quiz is to answer to 10 randomly chosen questions relative to the fields of molecular biology, biochemistry and genetics. Each month, we will randomly pick up one person among all those that have obtained a perfect score (and it's not so easy !) and will send that person some delicious **Swiss chocolate** !

Links have been added from SWISS-PROT to the Saccharomyces genomic database (SacchDb) at

Stanford.

A new SWISS-PROT document has been added, it is a nomenclature and index of allergen sequences.

May 26, 1995

A new service is available: SWISS-2DSERVICE. The Two-Dimensional Gel Electrophoresis Laboratory of Geneva, Switzerland, is running a highly reproducible method for the two-dimensional separation of proteins. The laboratory now provides a 2-D PAGE service to which you may send your samples for analysis. This service includes analytical and preparative high-resolution 2-D PAGE, electrotransfer on membranes and/or amino acid composition.

May 17, 1995

New link in the Tools page to the multiple sequence alignment at Washington University.

May 11, 1995

Two links have been added to the SWISS-PROT entries. The first one directly submits a request to Swiss-Model for a 3D model of the current SWISS-PROT protein. The result is then sent back by e-mail. The second one allows to perform a sequence alignment with the current sequence, using NCBI's Basic Local Alignment Search Tool. This link is especially interesting in the virtual SWISS-PROT entries produced by the Translate tool.

May 5, 1995

We announce a new service, SWISS-FLASH, that reports news of databases, software and services developments from the Swiss biocomputing groups responsible for the ECD, ENZYME, LISTA, PROSITE, SeqAnalRef, SWISS-2DPAGE, SWISS-3DIMAGE and SWISS-PROT databases; the Melanie software package; the WWW ExPASy server; the SWISS-Model, SWISS- Shop and other network-based computational tools; and the SWISS-2DSERVICE services. If you subscribe to this service, you will automatically get the SWISS-Flash bulletins by electronic mail.

The SWISS-3DIMAGE database has been completely reorganised and indexed. The database is now searchable in the same way as the other SWISS-*** databases. We now also supply pictures in JPEG format, in addition to GIF and SGI. The images may still be downloaded by FTP.

Links to REBASE points now the version maintained at John Hopkins, whose layout is nicer than our own text based version !

April 19, 1995

We added Translate, a new tool which allows the translation of a nucleotide (DNA/RNA) sequence to a protein sequence.

Most of the pages in the server have been "refreshed" to make them more readable.

March 21, 1995

Links have been added from SWISS-PROT to the LISTA database of budding yeast (Saccharomyces cerevisiae) genes coding for proteins prepared under the supervisation of Patrick Linder.

March 7, 1995

Links have been added from SWISS-PROT to the HSSP database of structure-sequence alignments from the Protein Design Group, EMBL, Heidelberg.

March 2, 1995

During the last two months, various links have been added:

- from SWISS-PROT to the SubtiList and YEPD databases
- from ENZYME to PROSITE and to the Ligand database in Kyoto
- internally from PROSITE entries to other relevant PROSITE entries

Links from SWISS-PROT to FlyBase use the new WWW server for that database.

Many new SWISS-PROT documents have been added.

The page on the Melanie 2-D PAGE analysis software has been completely redesigned and includes now a on-line tutorial, as well as a request for information form.

December 7, 1994

In order to help users navigate through the ExPASy server, we have added graphical examples. More will be added in the future. See for example: Celegans examples or the who's who on ExPASy page. Thanks to Brigitte Boeckmann for the illustrations.

October 31, 1994

ENZYME: the *ENZYME Data Bank* has been added to the ExPASy server. This database may be accessed by EC number, name, compound, cofactor, comment, or by browsing through the list of classes, subclasses and sub-subclasses. Any entry in SWISS-PROT that contains an EC number in the DE line has also a direct link to ENZYME (by clicking on the EC number).

October 20, 1994

New services:

- Swiss-Shop - a sequence allerting system for Swiss-Prot that allows you to automatically obtain new sequence entries relevant to your field(s) of interest.
- Swiss-Model - an automated knowledge-based protein modelling server.

Compute pI/Mw: the tool to compute pI and Mw now accepts also a list of ID/AC's.

**SWISS-PROT**: in PDB cross-reference lines, there is now a link called RASMOL, sending the PDB entry as a *chemical / pdb* MIME type. On Unix systems, if you add, in the file .mailcap in your home directory, a line of the form

```
chemical/pdb; rasmol %s
```

then RASMOL will automatically be launched to display the protein 3D structure. This works also with any other program which accepts PDB coordinates. On systems other than Unix, this may also be specified. See your browser's manual.

October 13, 1994

The SWISS-PROT top page has been re-modeled. A number of new functionalities and documents have been added.

October 7, 1994

New tools have been added:

- **Amino acid composition similarity search** - the search may now also be performed from a given SWISS-PROT entry, whose amino acid composition will be compared with the whole SWISS-PROT database.
- **Compute pI/Mw** - Compute the theoretical pI and Mw from a SWISS-PROT ID or AC, or for a given sequence.

October 5, 1994

The gels run during the 2-D PAGE courses in Geneva are now displayed on the server.

September 29, 1994

SWISS-2DPAGE: protein maps now have a pI/Mw scale.

SeqAnalRef: the *Sequence Analysis Bibliographic Reference* database has been added to the ExPASy server. This database may be accessed by keyword, by reference identifier, by author and by full text search.

List of on-line experts: in SWISS-PROT and PROSITE top pages, a list of on-line experts gives you the possibility to directly send questions to any of the listed experts. The list is ogranized by subjects.

SWISS-PROT: new lists added:

- List of abbreviations for journals cited
- List of species has been made active
- Yeast Chromosome III entries in SWISS-PROT
- Nomenclature of extracellular domain
- List of on-line experts

PROSITE: new 3D line with active links to PDB.

September 26, 1994

In the tool AACompIdent for identifying a protein by its amino acid composition, options have been added. They allow to specify how many proteins should be displayed, as well as the pI and Mw range in which the search should be performed.

Also, some old bugs have now been corrected.

September 12, 1994

The tool AACompIdent for identifying a protein by its amino acid composition, has been corrected and is now supposed to work. If you still encounter problems, please send us a mail.

June 17, 1994

SWISS-PROT: added cross-references (DR lines) to **GenBank**.

June 16, 1994

SWISS-PROT: added cross-references (DR lines) to **MaizeDB** Maize Genome Database of the National Agricultural Library.

June 6, 1994

Added the PROSITE page: PROSITE entries may now be searched by description of sites and pattern, by accession number, by author, and soon by full text search.

June 3, 1994

Added the GuessProt tool to the tools page: you may now get the SWISS-PROT proteins closest to a given *pI* and *Mw*.

May 27, 1994

In SWISS-PROT entries, added links to **GCRDb** - the *G-Protein--Coupled Receptor DataBase* .

Added the **list of nomenclature related references for proteins** to the SWISS-PROT top page.

May 26, 1994

Added a new reference 2-D PAGE map of **Platelet** to SWISS-2DPAGE.

May 20, 1994

The SWISS-2DPAGE team is now organizing a **2-D PAGE training** in Geneva once every three months.

May 18, 1994

Added the **Yeast Chromosome XI** list of proteins to the SWISS-PROT documentation page.

May 11, 1994

**Tools:** new page giving access to on-line analysis tools, such as BLAST, BLITZ, PROSITE search and amino acid composition analysis, and more to come in the future.

March 23, 1994

Added the list of **restriction enzymes and methylases** in SWISS-PROT top page.

March 22, 1994

The ExPASy WWW server has been upgraded to a **SPARCServer 10/51**. It should perform much faster now. If some features are not working, please tell us about.

March 18, 1994

The links to OMIM are now direct links to the **OMIM hypertext** server from GDB. Thanks to Keith Robison for informing me about it.

March 4, 1994

SWISS-2DPAGE: Added experimental **Amino Acid Composition Similarity Search** : you enter a protein's amino acid composition and the server will e-mail you the list of SWISS-PROT entries with similar compositions, sorted by decreasing similarity measure.

March 2, 1994

Added direct link to NCBI's **BLAST** Basic Local Alignment Search Tool (ExPASy and SWISS-PROT top pages).

March 1, 1994

Starting with release 28, **SWISS-PROT keyword search** will be performed on the main release as well as on the **weekly updates**.

In the SWISS-PROT page, added links to four additional active lists:

- Index of Escherichia coli K12 chromosomal entries in SWISS-PROT and their corresponding EcoGene cross-reference
- Index of Saccharomyces cerevisiae entries in SWISS-PROT and their corresponding gene designations
- Index of Caenorhabditis elegans entries in SWISS-PROT and their corresponding gene designations and WormPep cross-references
- Index of Dictyostelium discoideum entries in SWISS-PROT and their corresponding gene designations and DictyDB cross-references

February 23, 1994

Added two new reference 2-D PAGE maps: **Macrophage Like Cell Line** (U937) and **Erythroleukemia Cell** (ELC).

In a SWISS-2DPAGE entry, it is now possible to compute the **theoretical pI and Mw** of the protein.

February 14, 1994

Added **SWISS-2DPAGE Map Selection** : you select a 2-D PAGE reference gel, click on a spot and get information on the corresponding protein. See the SWISS-2DPAGE top page.

February 11, 1994

Added a new reference 2-D PAGE map of **Cerebrospinal Fluid** to SWISS-2DPAGE.

January 28, 1994

Added the **bionet newsgroups**.

January 25, 1994

Added an entry to **SWISS-3DIMAGE** images of crystallized proteins.

In SWISS-PROT entries which contain cross-references to PDB, added a cross-reference to **SWISS-3DIMAGE**. Try for example *AAT_ECOLI*.

January 24, 1994

Added **full text search** of the SWISS-PROT protein sequence database.

January 17, 1994

Added links to **MEDLINE** entries in SWISS-PROT, through NCBI's Entrez Server.

Added, in the SWISS-2DPAGE page, a link to the **QUEST Protein Database Center**.

December 1, 1993

Added a **User Survey**. Please help us inprove the server in participating to this survey.

Added a new reference 2-D PAGE map of **Lymphoma** to SWISS-2DPAGE.

November 23, 1993

Added link to **BioBit 24, the BIO-NAUT Newsletter** from November, 22, 1993, describing the World Wide Web.

November 18, 1993

Added links to the **Maize Genome Database** at Columbia, Missouri and to **EMBnet Switzerland**.

November 17, 1993

Added the list of overall **Top Ten** users in the *ExPASy server Activity Reports* page.

November 16, 1993

Added **Images of crystallized proteins** from this server.

Added links to **Harvard Biological Laboratories**, the **Gene-Server** at University of Houston, the **EMBnet**: Biocomputing in Europe, the **biology servers index** at USGS, **Jackson Laboratory** WWW server and Keith Robison's **Molecular Biology WWW sampler**.

October 12, 1993

Added a list of **specialised documents** to the SWISS-PROT top page, such as 7-transmembrane G-linked receptors, CD nomenclature for surface proteins of Human leucocytes and Vertebrate homeobox proteins. Some of these list give then direct access to corresponding SWISS-PROT entries.

October 8, 1993

Added links to the **Caenorhabditis elegans** and **Mycobacterium** databases at INRA (France).

Added a link to the **ExPASy server activity** reports.

October 4, 1993

Moved to the NCSA server.

September 28, 1993

Added the **PDB** Brookhaven Protein Data Bank of 3D structures. In SWISS-PROT, cross-references to PDB have now active links to the gopher server at Protein Data Bank. You may access the PDB entry or get the 3D image. Try for example the SWISS-PROT entry *P00782*.

September 27, 1993

Added the **FlyBase** database of genetic and molecular data for Drosophila. In SWISS-PROT and EMBL, cross-references to FLYBASE are now active links. Therefore, SWISS-PROT has now active links to SWISS-2DPAGE, EMBL, PROSITE, REBASE, OMIM and FLYBASE. EMBL has active links to SWISS-PROT and FlyBase.

September 23, 1993

Added a link to the National Institute of Health **Genobase** server to our top page.

September 21, 1993

Announced the ExPASy server and SWISS-2DPAGE release 0 to bionet.announce.

August 1, 1993

Installed the ExPASy molecular biology server, release 0, beta version.

---

*Last modified 21/Oct/2004 by CHH*